Students' Mathematics Self-Efficacy: Relationship with Test Achievement and Development in the Classroom

Karin Elisabeth Sørlie Street
Department of Education
University of Oxford
St Cross College



Supervisors
Dr Lars-Erik Malmberg
Professor Gabriel Stylianides

Thesis submitted to the University of Oxford in fulfilment of the requirements for the degree of Doctor of Philosophy in Education

Hilary Term, 2018

Abstract

Self-efficacy is individuals' judgments regarding their capabilities to carry out future tasks or challenges. These judgments of capability are related to important learning behaviours such as effort and persistence, performance, and choice of career path. In order to support students' continued engagement with and learning of mathematics, it is important to consider how students make sense of their mathematical experiences as well as the relationship between students' mastery experiences and mathematics self-efficacy. In this thesis I address important gaps in the literature in regard to the conceptualisation of the self-efficacy construct, the relationship between self-efficacy and mathematics performance, the stability and change of self-efficacy when learning new topics in mathematics, and self-efficacy development over a series of lessons in mathematics across cycles of self-efficacy and mastery experiences.

The thesis included two phases of data collection and analysis. The first phase involved students in grades 5, 8, and 9 (N = 756) and included measures of students' self-efficacy and national test performance. The second phase involved students in grades 6 and 10 (N = 181) and included repeated measures of students' self-efficacy and mastery experiences from a series of lessons in mathematics, when students were introduced to new topics. I analysed the data using different methods, including confirmatory factor analyses to investigate the structural validity of my measures, and structural equation models to investigate stability and change over time, and relationships between constructs.

Major findings from the analysis include the following: students considered levels of difficulty when appraising their experiences in mathematics and when forming their self-efficacy; students' test performance predicted their self-efficacy one year later, but not the other way around; the mean-level of students' self-efficacy grew significantly across lessons when students were introduced to new topics, even as the rank-order of their self-efficacy remained highly stable; and there was a reciprocal relationship between students' self-efficacy and their mastery experiences, where substantial effects from both constructs on gains in the other construct remained stable across a sequence of lessons in mathematics.

The findings have important implications for how we conceptualise self-efficacy, mastery experiences, and their relationship over time. Furthermore, the findings from my thesis have implications for teacher practice. In order to support adaptive self-efficacy, teachers need to consider the experiences students have with mathematics, not just the skills they learn. If teachers themselves gain knowledge about how individual students make sense of their mathematical experiences, they can support students' appraisals of these experiences, and prevent maladaptive cycles from occurring. In short, students need support not just to develop their mathematical skills, but also to develop adaptive appraisals of their mathematical experiences, in order to form self-efficacy beliefs that are reflective of each student's potential to learn mathematics.

Acknowledgements

I would like to express my sincere gratitude to my supervisors Lars-Erik Malmberg and Gabriel Stylianides. I would like to thank Lars, for teaching me about structural equation models, and for always being enthusiastic about my data. I would like to thank Gabriel, for introducing me to the field of mathematics education, and for setting a new standard for concise writing. Beyond giving patient and constructive feedback on my work, they have also known when to provide encouragement and support. I feel fortunate for having had the opportunity to work with, and learn from, such inspiring mentors.

Furthermore, I am grateful to my examiners for the transfer of status and confirmation reports: Pam Sammons, Therese Hopfenbeck, and Jo-Anne Baird, for their constructive feedback and useful comments on earlier stages of my work.

I owe a debt of gratitude to my Norwegian institution, Høgskulen i Sogn & Fjordane, for granting me the funds to undertake this DPhil in the first place, and to St Cross College, for support during my time in the UK.

I am also thankful to all my colleagues, fellow students, and friends, in Norway and in the UK, who have provided engaging discussions, study-dates, fresh-air adventures, and IT support.

To all my family, who have been instrumental in providing motivation, baby-sitting, accommodation, hot tea, conversations, chocolate, and so much more… You are wonderful, thank you!

Finally, to my dearest Alistair and Klara, who have provided boundless enthusiasm, and made sure I remember the important things in life. Thank you for your endless support and love.

Ali & Klara, I dedicate this thesis to you.

List of content

1 INTRODUCTION 13
1.1 MAIN THEMES AND WHY THEY ARE IMPORTANT 14
1.2 STRUCTURE OF THE THESIS 16
1.2.1 Level, strength, and facet-specific self-efficacy (study 1) 17
1.2.2 Reciprocal relationships between self-efficacy and national test performance (study 2) 19
1.2.3 The stability and change of students' self-efficacy when learning a new topic (study 3) 19
1.2.4 The relationship between self-efficacy and mastery experiences across a sequence of lessons (study 4) 20
1.2.5 Overall discussion and conclusion 21
1.3 BRIEF THEORETICAL BACKGROUND 21
1.3.1 Conceptual underpinnings of self-efficacy 22
1.3.2 Relationship between self-efficacy and performance 24
1.3.3 Development and change of self-efficacy 27
1.4 A MODEL OF STUDENTS' SELF-EFFICACY IN MATHEMATICS 31
1.5 CONTRIBUTION TO KNOWLEDGE 34
1.5.1 Research questions 36

2 LEVEL, STRENGTH, AND FACET-SPECIFIC SELF-EFFICACY, AND STUDENTS' PERFORMANCE ON NATIONAL TESTS IN MATHEMATICS (STUDY 1) 38
2.1 INTRODUCTION 38
2.2 THEORETICAL BACKGROUND 40
2.2.1 Facet-specificity 43
2.2.2 Strength 52
2.2.3 Level 57
2.2.4 Theoretical model and research questions 61
2.3 METHODOLOGY 64
2.3.1 Sample 64
2.3.2 Ethical considerations 67
2.3.3 Measures 70
2.3.4 Empirical models 85
2.3.5 Analyses 93
2.4 RESULTS 99
2.4.1 Correlation coefficients 99
2.4.2 RQ1: Confirmatory factor analyses 104
2.4.3 RQ2: Associations between national test results and self-efficacy 113
2.4.4 RQ3: Self-efficacy and national test scores for easy, medium, and hard tasks 121
2.5 DISCUSSION 131
2.5.1 RQ1: Structural validity of the proposed measure 131
2.5.2 RQ2: Multidimensional self-efficacy and national test scores in mathematics 133
2.5.3 RQ3: Self-efficacy constructs matched with national test parcels 135
2.5.4 Limitations 142
2.6 CONCLUSION 144

3 RECIPROCAL RELATIONSHIPS BETWEEN SELF-EFFICACY AND NATIONAL TEST PERFORMANCE (STUDY 2) 146
3.1 INTRODUCTION 146
3.2 THEORETICAL BACKGROUND 147
3.2.1 Theoretical model and research questions 151
3.3 METHODOLOGY 153

3.3.1 Participants ... *153*
3.3.2 Measures ... *154*
3.3.3 Specification of empirical models ... *156*
3.3.4 Analyses ... *158*
3.4 RESULTS ... 158
3.4.1 Correlation coefficients ... *158*
3.4.2 Structural equation models ... *161*
3.5 DISCUSSION ... 164
3.5.1 RQ1: Reciprocal effects ... *165*
3.5.2 RQ2: Differential relationships between self-efficacy levels and facets, and performance results ... *167*
3.6 CONCLUSION ... 168

4 STABILITY AND CHANGE OF STUDENTS' SELF-EFFICACY WHEN LEARNING A NEW TOPIC (STUDY 3) ... 170
4.1 INTRODUCTION ... 170
4.2 THEORETICAL BACKGROUND ... 173
4.2.1 Changes across several years ... *175*
4.2.2 Changes across a single year ... *179*
4.2.3 Changes across learning events ... *183*
4.2.4 Summary of research on self-efficacy change ... *184*
4.2.5 Theoretical model and research questions ... *188*
4.3 METHODOLOGY ... 190
4.3.1 Overall design ... *191*
4.3.2 Participants ... *193*
4.3.3 Procedure ... *200*
4.3.4 Further ethical considerations ... *202*
4.3.5 Measures ... *203*
4.3.6 Analyses ... *214*
4.4 RESULTS ... 226
4.4.1 Reliable change statistics ... *227*
4.4.2 RQ1: The structural validity of my lesson self-efficacy measure ... *232*
4.4.3 RQ2: Stability of students' rank-order ... *251*
4.4.4 RQ3: Self-efficacy change across lessons ... *258*
4.5 DISCUSSION ... 289
4.5.1 Reliable changes in self-efficacy ... *289*
4.5.2 A multidimensional measure of self-efficacy for lessons in mathematics ... *290*
4.5.3 High rank-order stability of students' self-efficacy ... *292*
4.5.4 Mean-level changes in self-efficacy ... *296*
4.5.5 Further implications from study 3 ... *302*
4.5.6 Further limitations of study 3 ... *304*
4.6 CONCLUSION ... 305

5 THE RELATIONSHIP BETWEEN SELF-EFFICACY AND MASTERY EXPERIENCES ACROSS A SEQUENCE OF MATHEMATICS LESSONS (STUDY 4) ... 307
5.1 INTRODUCTION ... 308
5.2 THEORETICAL BACKGROUND ... 309
5.2.1 Reciprocal effects between mathematics self-efficacy and mastery experiences ... *314*
5.2.2 Reciprocal effects in related domains ... *319*
5.2.3 Theoretical model and research questions ... *322*
5.3 METHODOLOGY ... 325
5.3.1 Design ... *325*
5.3.2 Measures ... *327*
5.3.3 Analyses ... *339*
5.4 RESULTS ... 345

5.4.1 RQ1: The structural validity of my mastery experiences measure *346*
5.4.2 RQ2: The relationship between students' self-efficacy and mastery experiences over time *359*
5.4.3 RQ3: The relationship between students' self-efficacy, mastery experiences, and ability conceptions *384*
5.5 DISCUSSION 412
5.5.1 A multidimensional measure of mastery experiences from lessons in mathematics .. *412*
5.5.2 Reciprocal relationships between self-efficacy and mastery experiences *413*
5.5.3 Significant relationships with ability conceptions *424*
5.5.4 Further implications and limitations of my study *430*
5.6 CONCLUSIONS 434

6 OVERALL DISCUSSION AND CONCLUSION 435
6.1 CROSS-STUDY FINDINGS AND IMPLICATIONS 440
6.1.1 The conceptualisation of self-efficacy (and mastery experiences) *440*
6.1.2 The relationship between self-efficacy and performance *442*
6.1.3 The development and change of self-efficacy *445*
6.2 LIMITATIONS 448
6.3 CONCLUSION 451

References 456

Appendices 469

List of Figures

Introduction

Figure 1: Theoretical model for the thesis 32

Study 1

Figure 2: Theoretical model for study 1 63
Figure 3: Unidimensional (Model 1) 88
Figure 4: 4 correlated latent facets (Model 2) 89
Figure 5: 4 correlated latent facets, with CUs* for 3 levels (Model 2b) 89
Figure 6: 3 correlated latent levels (Model 3) 89
Figure 7: 3 correlated latent levels, with CUs* for 4 facets (Model 3b) 90
Figure 8: 2nd-order construct with 4 latent facets (Model 4) 90
Figure 9: 2nd-order construct with 4 latent facets & CUs* for 3 latent levels (Model 4b) 91
Figure 10: 2nd-order construct with 3 latent levels (Model 5) 91
Figure 11: 2nd-order construct with 3 latent levels & CUs* for 4 facets (Model 5b) 92
Figure 12: "LMF" model with 4 latent facets & 3 latent levels (Model 6) 92
Figure 13: 2nd -order "LMF" model with 4 latent facets & 3 latent levels (Model 7) 93
Figure 14: Structural equation model (Model 3b) 119
Figure 15: Model with easy, medium, & hard national tests on self-efficacy, grade 5 127
Figure 16: Model with easy, medium, & hard national tests on self-efficacy, grade 8 128
Figure 17: Model with easy, medium, & hard national tests on self-efficacy, grade 9 129

Study 2

Figure 18: Theoretical model for study 2 152
Figure 19: Structural equation model, self-efficacy level constructs (Model 1) 162
Figure 20: Structural equation model, self-efficacy facet constructs (Model 2) 163

Study 3

Figure 21: Overview of empirical findings on Self-Efficacy change 187
Figure 22: Theoretical model for study 3 189
Figure 23: Design of phase 2 192
Figure 24: Model 0, Unidimensional 219
Figure 25: Model 1, Latent Facets 220
Figure 26: Model 2, Latent Levels 220
Figure 27: Model 3, Bi-factor Easy 221
Figure 28: Model 4, Bi-factor Medium difficulty 221
Figure 29: Model 5, Bi-factor Hard 222
Figure 30: Latent Levels model, lesson 2 244
Figure 31: Invariance model self-efficacy, easy level construct 249
Figure 32: Invariance model self-efficacy, medium difficulty level construct 249
Figure 33: Invariance model self-efficacy, hard level construct 250
Figure 34: Autoregressive model Self-Efficacy, easy tasks 255
Figure 35: Autoregressive model Self-Efficacy, medium difficulty tasks 256
Figure 36: Autoregressive model Self-Efficacy, hard tasks 257
Figure 37: Linear growth model Self-Efficacy easy tasks, grade 6 268
Figure 38: Unconditional growth model Self-Efficacy easy tasks, grade 10 269
Figure 39: Estimated growth curves, Self-Efficacy easy tasks 269
Figure 40: Quadratic growth model medium difficulty tasks, grade 6 274
Figure 41: Unconditional growth model medium difficulty tasks, grade 10 274

Figure 42: Estimated growth curves, Self-Efficacy medium difficulty tasks 275
Figure 43: Quadratic growth model hard tasks, grade 6 279
Figure 44: Unconditional growth model hard tasks, grade 10 280
Figure 45: Estimated growth curves, Self-Efficacy hard tasks 280
Figure 46: Estimated growth curves for Self-Efficacy, grade 6 and grade 10 288

Study 4

Figure 47: Theoretical model for study 4 323
Figure 48: Design of study 4 326
Figure 49: Unidimensional Model (Model 0) 341
Figure 50: Latent Facets Model (Model 1) 341
Figure 51: Latent Levels Model (Model 2) 342
Figure 52: Bi-factor Model, easy tasks (Model 3) 342
Figure 53: Bi-factor Model, medium difficult tasks (Model 4) 343
Figure 54: Bi-factor Model, hard tasks (Model 5) 343
Figure 55: Latent Levels model, lesson 1 354
Figure 56: Invariance model mastery experiences, easy construct 355
Figure 57: Invariance model mastery experiences, medium difficulty construct 356
Figure 58: Invariance model mastery experiences, hard construct 356
Figure 59: Autoregressive model Mastery Experiences, easy tasks 367
Figure 60: Autoregressive model Mastery Experiences, medium difficulty tasks 368
Figure 61: Autoregressive model Mastery Experiences, hard tasks 369
Figure 62: Autoregressive path model of Task Self-Efficacy 375
Figure 63: Cross-lagged path model of Self-Efficacy and Mastery Experiences, easy level constructs 379
Figure 64: Cross-lagged path model of Self-Efficacy and Mastery Experiences, medium difficulty level constructs 380
Figure 65: Cross-lagged path model of Self-Efficacy and Mastery Experiences, hard level constructs 381
Figure 66: Simplex model of Self-Efficacy, easy tasks, with Ability Conceptions 396
Figure 67: Simples model of Self-Efficacy, medium difficulty tasks, with Ability Conceptions 397
Figure 68: Simplex model of Self-Efficacy, hard tasks, with Ability Conceptions 398
Figure 69: Simplex model of Mastery Experiences, easy tasks, with Ability Conceptions 401
Figure 70: Simplex model of Mastery Experiences, medium difficulty tasks, with Ability Conceptions 402
Figure 71: Simplex model of Mastery Experiences, hard tasks, with Ability Conceptions 403
Figure 72: Cross-lagged path model of Self-efficacy and Mastery Experiences, easy tasks, with Ability Conceptions 407
Figure 73: Cross-lagged path model of Self-efficacy and Mastery Experiences, medium difficulty tasks, with Ability Conceptions 408
Figure 74: Cross-lagged path model of Self-efficacy and Mastery Experiences, hard tasks, with Ability Conceptions 409

List of Tables

Study 1

Table 1: Participants study 1 65
Table 2: Item descriptives, grade 5 75
Table 3: Item descriptives, grades 8 and 9 76
Table 4: Comparison of national test performance (grade 5, 2012) 79
Table 5: Comparison of national test performance (grade 5, 2013) 80
Table 6: Comparison of national test performance (grade 8, 2012) 81
Table 7: Comparison of national test performance (grade 9, 2013) 83
Table 8: Zero-order correlation coefficients among items 101
Table 9: Zero-order correlation coefficients among latent variables, grade 5 102
Table 10: Zero-order correlation coefficients among latent variables, grade 8 103
Table 11: Zero-order correlation coefficients among latent variables, grade 9 103
Table 12: Results from confirmatory factor analyses, grade 5 105
Table 13: Δ fit indices grade 5 models* 105
Table 14: Results from confirmatory factor analyses, grades 8 & 9 109
Table 15: Δ fit indices grade 8 & 9 models* 109
Table 16: Results from structural equation models, grade 5 114
Table 17: Δ fit indices grade 5 structural equation models* 114
Table 18: Results from structural equation models, grades 8 & 9 117
Table 19: Δ fit indices grades 8 and 9 structural equation models* 117
Table 20: Accuracy and bias, grade 5 121
Table 21: Accuracy and bias, grade 8 122
Table 22: Accuracy and bias, grade 9 124
Table 23: Results from path models with matched self-efficacy and test items 126

Study 2

Table 24: Item descriptives 155
Table 25: Correlation matrix of self-efficacy grade 8 and grade 9 160
Table 26: Fit indices from structural equation models 161

Study 3

Table 27: Phase 2 participants at each time point 194
Table 28: Topic and main tasks for each class / lesson 195
Table 29: Phase 1 test-scores, self-efficacy, and correlations, grade 5 classes 197
Table 30: Phase 1 test-scores, self-efficacy, and correlations, grade 8 & 9 classes 198
Table 31: Descriptives Self-Efficacy pre-observations 206
Table 32: Descriptives Self-Efficacy lesson 1 209
Table 33: Descriptives Self-Efficacy lesson 2 210
Table 34: Descriptives Self-Efficacy lesson 3 211
Table 35: Descriptives Self-Efficacy lesson 4 212
Table 36: Self-Efficacy mean, lesson by lesson 213
Table 37: Correlation matrix Self-Efficacy level constructs, pre-obs. - lesson 4 228
Table 38: Reliable increase, decrease or no change in Self-Efficacy 230
Table 39: Reliable change in Self-Efficacy across all time points 231
Table 40: Correlation matrix Self-efficacy, pre-observations 234
Table 41: Correlation matrix Self-efficacy, lesson 1 235
Table 42: Correlation matrix Self-efficacy, lesson 2 235
Table 43: Correlation matrix Self-efficacy, lesson 3 236

Table 44: Correlation matrix Self-efficacy, lesson 4 236
Table 45: Fit indices from confirmatory factor analyses, Self-Efficacy pre-obs 239
Table 46: Fit indices from confirmatory factor analyses, Self-Efficacy lesson 1 240
Table 47: Fit indices from confirmatory factor analyses, Self-Efficacy lesson 2 241
Table 48: Fit indices from confirmatory factor analyses, Self-Efficacy lesson 3 242
Table 49: Fit indices from confirmatory factor analyses, lesson 4 243
Table 50: Fit indices Self-Efficacy invariance testing 247
Table 51: Fit indices Self-Efficacy simplex models 252
Table 52: Variance component models, Self-Efficacy easy level constructs 260
Table 53: Variance comp. models, Self-Efficacy medium difficulty level constructs 260
Table 54: Variance component models, Self-Efficacy hard level constructs 260
Table 55: Difference from average class, self-efficacy easy tasks 261
Table 56: Difference from average class, self-efficacy medium difficulty tasks 261
Table 57: Difference from average class, self-efficacy hard tasks 262
Table 58: Fit indices growth models easy tasks, grade 6 265
Table 59: Parameter estimates easy task models, grade 6 265
Table 60: Fit indices growth models easy tasks, grade 10 266
Table 61: Parameter estimates easy task models, grade 10 266
Table 62: Fit indices growth models medium difficulty tasks, grade 6 271
Table 63: Parameter estimates medium difficulty task models, grade 6 271
Table 64: Fit indices growth models medium difficulty tasks, grade 10 272
Table 65: Parameter estimates medium difficulty task models, grade 10 272
Table 66: Fit indices growth models hard tasks, grade 6 277
Table 67: Parameter estimates hard task models, grade 6 277
Table 68: Fit indices growth models hard tasks, grade 10 278
Table 69: Parameter estimates hard task models, grade 10 278
Table 70: Group models with equality constraints, easy tasks 283
Table 71: Group models with equality constraints, medium difficulty tasks 283
Table 72: Group models with equality constraints, hard tasks 284

Study 4

Table 73: Descriptives for the Task Self-Efficacy measure 329
Table 74: Descriptives Mastery Experiences, lesson 1 332
Table 75: Descriptives Mastery Experiences, lesson 2 333
Table 76: Descriptives Mastery Experiences, lesson 3 334
Table 77: Descriptives Mastery Experiences, lesson 4 335
Table 78: Mastery Experiences mean scores, lesson by lesson 336
Table 79: Descriptives Ability Conceptions 338
Table 80: Correlation matrix Mastery Experiences, lesson 1 347
Table 81: Correlation matrix Mastery Experiences, lesson 2 347
Table 82: Correlation matrix Mastery Experiences, lesson 3 348
Table 83: Correlation matrix Mastery Experiences, lesson 4 348
Table 84: Results confirmatory factor analyses, Mastery Experiences lesson 1 350
Table 85: Results confirmatory factor analyses, Mastery Experiences lesson 2 351
Table 86: Results confirmatory factor analyses, Mastery Experiences lesson 3 352
Table 87: Results confirmatory factor analyses, Mastery Experiences lesson 4 353
Table 88: Fit indices Mastery Experiences invariance tests 358
Table 89: Correlation matrix Mastery Experiences level constructs, lesson 1-4 362
Table 90: Fit indices Mastery Experiences simplex models 365

Table 91: Autocorrelations for the single-item Task Self-Efficacy measure 370
Table 92: Fit indices Task Self-Efficacy simplex models 372
Table 93: Fit indices cross-lagged path models with equality constraints 378
Table 94: Correlation matrix Ability Conceptions 387
Table 95: Results confirmatory factor analysis, Ability Conceptions 389
Table 96: Correlation matrix Ability Conceptions, Self-efficacy, and Mastery Experiences, lessons 1-4 391
Table 97: Results Self-Efficacy models with Ability Conceptions as covariates 395
Table 98: Results Mastery Experiences models w/ Ability Conceptions as covariates 400
Table 99: Results Task Self-Efficacy model with Ability Conceptions as covariates 405
Table 100: Fit indices of path models of Self-Efficacy, Mastery Experiences, Task Self-Efficacy, with Ability Conceptions as covariates 406
Table 101: Overview of research questions and findings from my thesis 437

List of Appendices

Appendix A: Article: Street, Malmberg & Stylianides, 2017a…….. 469
Appendix B: Conference proceedings: Street, Malmberg & Stylianides, 2017b……...498
Appendix C: Differences between thesis chapter and a published article.…….506
Appendix D: Participant information letter, students, phase 1…….……..508
Appendix E: Participant information letter, teachers /schools, phase 1……. 510
Appendix F: NSD & CUREC approval ……...……..512
Appendix G: Item descriptives, including 5th "grade" facet …..…….....514
Appendix H: Grade 8/9 corr. matrix, including 5th "grade" facet …….…….....515
Appendix I: Self-Efficacy Gradations of Difficulty (SEGD), test version…….……516
Appendix J: Self-Efficacy Gradations of Difficulty (SEGD), classroom version……..517
Appendix K: Self-Efficacy Gradations of Difficulty (SEGD), short classroom version.....519
Appendix L: Participant information letter, students, phase 2…….……...520
Appendix M: Participant information letter, teachers /schools, phase 2……..522
Appendix N: Mastery Experiences Gradations of Difficulty (MEGD)……...524
Appendix O: Self-Efficacy Task instrument ……. ..…….525
Appendix P: Ability Conceptions Questionnaire……...……..……526
Appendix Q: Instruments in Norwegian: Student questionnaires, phase 1 ……..…...527
Appendix R: Instruments in Norwegian: Student questionnaires, phase 2 …..…..…..…...530

1 Introduction

What is the relationship between the kinds of experiences students have with mathematics, and their confidence that they can perform well on mathematical tasks? Further, how does students' confidence for future learning and performance in mathematics develop during classroom lessons? These overarching questions guide my investigations in the present thesis. To address these questions, I will focus on students' self-efficacy in mathematics, and the relationship between students' self-efficacy and their mathematical experiences, such as taking national tests or working on mathematical tasks. I will also focus on how we can measure self-efficacy, how we can measure students' mastery experiences with mathematics, and how self-efficacy develop across classroom lessons while learning mathematics. Students' performance in mathematics is important as it is an entry-requirement into a range of career fields, and thus critical to the career options available to students (Lent, Lopez, Brown, & Gore Jr., 1996). Students' beliefs and attitudes about mathematics influence both performance and choices to engage with the subject or pursue mathematics related careers (Lent, Lopez, & Bieschke, 1993; Muis, 2004; Multon, Brown, & Lent, 1991; Zeldin & Pajares, 2000).

While the main themes of my thesis, which I discuss in the next section, are not limited to specific countries, the study is set in Norway, where researchers have demonstrated a decline in students' motivation between 5th and 10th grade (Skaalvik & Skaalvik, 2009). Several researchers have found that Norwegian classrooms score high on emotional support, but are lacking in instructional support to students (Gamlem & Munthe, 2014; Klette, 2011). Some claim mathematics teaching in Norwegian schools is traditional and does not support students as active learners (Klette et al., 2008). Knowledge is needed regarding the relationship

between motivational variables and achievement outcomes in mathematics in Norwegian schools, as well as how to increase student motivation in the classroom.

In this chapter, I will present the themes of the current thesis in relation to the two overarching questions presented above. My thesis includes four interrelated studies, presented in separate chapters, where each study builds on and extends the themes from the proceeding studies. One challenge of the current chapter is to present the theoretical background relevant for the thesis, while also avoiding unnecessary repetition of the theoretical background that will be addressed in the chapter of each study. Furthermore, due to their interrelatedness, there is also overlap across my studies in relation to key theoretical assumptions. To address this challenge I will in the current chapter present only the brief and overarching theoretical background necessary to propose a theoretical model for the thesis, while I will address relevant theory more thoroughly within the chapter for each study. Furthermore, some repetition will occur across the studies, in regard to key theoretical tenets such as the conceptualisation of self-efficacy.

In the next section I will present the main themes for my thesis, and why they are important to consider. I will then provide an overview of the organization of the thesis and how the different chapters are related, before I present a brief theoretical background for my themes. At the end of the chapter I will discuss how I will contribute to knowledge in light of these themes, through addressing separate, but related, research questions.

1.1 Main themes and why they are important

The main theme of the current thesis is students' self-efficacy in mathematics. *Self-efficacy* is defined as individual beliefs people form regarding their own capabilities to perform particular tasks: "beliefs in one's capabilities to organize and execute the courses of action

required to produce given attainments" (Bandura, 1997, p. 3). Self-efficacy is important because numerous studies have linked self-efficacy with factors related to adaptive learning behaviours, motivation constructs, career choices and performance outcomes (see Bråten, Samuelstuen, & Strømsø, 2004; Hackett, 1995; Klassen & Usher, 2010; Pajares, 1996b; Schulz, 2005; Zimmerman, 2000).

Self-efficacy in mathematics is important because self-efficacy predicts students' choice of engaging with challenging tasks (Bandura & Schunk, 1981) and are related to strategies and types of achievement goals students set for learning mathematics (Bong, 2009; Throndsen, 2011; Zimmerman & Martinez-Pons, 1990), and students' mathematics problem-solving and performance (Hackett & Betz, 1989; McConney & Perry, 2010; Pajares, 1996a). Self-efficacy is also related to students' physiological states while thinking about and engaging with mathematical tasks (Matsui, Matsui, & Ohnishi, 1990; Schulz, 2005), their career choices (Hackett & Betz, 1989) and their future intent and disengagement with mathematics (Martin, Anderson, Bobis, Way, & Vellar, 2012; Pampaka, Kleanthous, Hutcheson, & Wake, 2011).

Students' experiences with mathematics are another important theme of this thesis. Students judge their experiences (e.g. on tests or during classroom learning) in association with their beliefs about effort and ability, where these appraised experiences are related to their confidence to perform on future tasks (Butz & Usher, 2015; Schweinle, Berg, & Sorenson, 2013). As such students' appraised performance experiences (e.g. mastery experiences) are important sources for future self-efficacy (Usher & Pajares, 2008).

Finally, the current study is focused on methodologies to study self-efficacy and students' experiences with mathematics. In order to contribute meaningfully to the literature about

students' self-efficacy, it is important to use instruments that are aligned with the dominant conceptualisation of the construct, such as the proposed multidimensional nature of self-efficacy (Bandura, 1997). Furthermore, there is a need for investigations into learning processes (Schmitz, 2006; Vermunt & Verloop, 1999), and methodologies that measure processes in education, for instance how students' self-efficacy beliefs develop in the classroom, in relation to students' experiences with mathematics.

The present study draws on knowledge from several fields of study. The specific view is taken of learners as active agents who continuously make sense of information from the environment in light of previous experiences and cognitions. This is in line with Bandura's social cognitive theory (1997) and Zimmerman's model for development of self-efficacy and self-regulation (Zimmerman, 1990, 2000). While learning in the current study is seen as an individual process, related to cognitive change, there is according to social-cognitive theory a reciprocal relationship between learners' cognitions, behaviours and their learning environment: they both influence and are being influenced by one another. Vermunt & Verloop defined learning as "self-regulated knowledge construction" (1999, p. 258), which serves to highlight the importance of motivational self-beliefs, as they relate to self-regulative behaviours (Mega, Ronconi, & De Beni, 2014). In a life-long learning perspective, where students should learn not just skills but also how to learn, self-regulation is central (Hopfenbeck, 2011; Throndsen, 2011). Before I elaborate on the overarching theoretical perspectives guiding this thesis, I will introduce the reader to how the thesis is structured.

1.2 Structure of the thesis

The present document is organised according to 11 research questions, which are investigated in four studies. The studies are stand-alone, but also related. Data for studies 1 and 2 was collected together, during the National Test Phase (phase 1), while the data collection for

studies 3 and 4 was coined the Classroom Development Phase (phase 2). Significantly shorter versions of studies 1 and 2 have been published in a mathematics education journal (Street, Malmberg, & Stylianides, 2017a) and in the proceedings of a European mathematics education conference (Street, Malmberg, & Stylianides, 2017b), respectively (see Appendices A & B).

In the chapter for each study I present an introduction to the themes of that study, discuss the theoretical background, my method for investigating the current themes, and present the results, before I discuss the results in view of theory and previous studies. A challenge with the structure of this document is to avoid replication of information, while also facilitating readability. Each study is related to the others through the main themes of the thesis, thus a reference to the conceptualisation of self-efficacy would be necessary in order for each study to be stand-alone. Furthermore, the studies are based on data stemming from the same data collections (conducted over two phases), thus certain parts of their methodology are shared. I have decided to include in each chapter a brief introduction about the construct of self-efficacy. This leads to some repeat information, but the readability would be impaired if this information was not included. With regard to methodology, I in each study present all information applying to that study alone, while I refer back to previous chapters for information on shared methodological aspects. For instance, in regard to ethical considerations in studies 2 and 4 I refer back to studies 1 and 3, respectively, as the participants and the design was shared for studies belonging to the same phase of data collection. I will now present the remaining chapters of the current document.

1.2.1 Level, strength, and facet-specific self-efficacy (study 1)

In this study I investigate the structural validity of a new instrument, and use it to investigate the relationship between self-efficacy and national test results. The instrument responds to the

dominant conceptualisation of self-efficacy in the literature as a multidimensional construct, through including level, strength, and facet-specific self-efficacy. In the study I present my conceptualisation of self-efficacy, before I review existing measures and how they have contributed to knowledge. I argue that while there are measures that take into account strength and specificity of self-efficacy, seemingly no previous measures of mathematics self-efficacy included levels of perceived difficulty. I present a theoretical model of the three-dimensional self-efficacy construct, before I test plausible empirical models using confirmatory factor analysis and structural equation modelling in Mplus. I use the best-fit model to investigate the relationship between level and facet-specific self-efficacy, and performance on national tests. Data for this study was collected from Norwegian students who participated during the National Test Phase (phase 1) of the data collection, which was designed prior to the start of my DPhil in Oxford, October 2013. Study 1 of my thesis is linked with a major project funded by the Norwegian research council (NFR), termed Learning Regions (Langfeldt, 2015). Study 1 and the "Mestringsforventninger" study (Sørlie & Söderlund, 2015) of the Learning Regions project use the same participants and data, however different research questions are pursued. I elaborate on the study 1 sample in section 2.3.1.

As mentioned above, some aspects of study 1 have been previously published in a journal article (Street et al., 2017a). However, there are important differences between the published article (Appendix A) and the chapter in my thesis. A discussion of these differences is included in Appendix C, which also includes a discussion of the authorship of the published article and my doctoral thesis.

1.2.2 Reciprocal relationships between self-efficacy and national test performance (study 2)

Study 2 builds on study 1, in that it employs the instrument introduced in study 1 to investigate a theoretically proposed reciprocal effects models between students' self-efficacy and their national test performances. I argue that while the relationship between self-efficacy and performance has been extensively tested, only few studies have investigated empirically reciprocal effects between self-efficacy and performance results in mathematics. Furthermore, it seems that no studies have investigated this relationship using a measure including level, strength, and facet-specific self-efficacy. I use the results from study 1 to test two models, parcelling self-efficacy items according to levels of difficulty or test-taking facets. Through this practical application of the newly developed instrument, separate estimates are produced for the relationship between level-specific and facet-specific self-efficacy respectively, and national test performance. The data for this study was collected during phase 1 of the data collection, and the participants are a subsample of the students from study 1, who filled out self-efficacy questionnaires and sat national tests in mathematics two years in a row. I presented some aspects of my work on study 2 at a mathematics education conference, and a small part of this study is reported in a short paper that appears in the conference proceedings (Street et al., 2017b; see Appendix B).

1.2.3 The stability and change of students' self-efficacy when learning a new topic (study 3)

Study 3 builds on theoretical and methodological ideas from studies 1 and 2. In this study I use repeated measures in a micro-longitudinal design, to investigate the development of students' self-efficacy in mathematics. While we have "snapshots" of students' self-efficacy through cross-sectional studies, and some studies indicate students' self-efficacy declines with age, I argue there is a need for more knowledge about how self-efficacy develops over a short

time frame, when students learn new mathematical topics. In order to investigate this, I adapt the multidimensional self-efficacy instrument from phase 1. I use a baseline measure for students' self-efficacy, and repeated measures of students' self-efficacy across a sequence of lessons in mathematics, as students were introduced to new topics in mathematics. I test the structural validity of my proposed self-efficacy measure, and investigate invariance of this measure over time. I then investigate stability and change of self-efficacy using autoregressive path models and latent growth models for change. The data for this study stems from the phase 2 data collection, which was designed together with my supervisors, Lars-Erik Malmberg and Gabriel Stylianides.

1.2.4 The relationship between self-efficacy and mastery experiences across a sequence of lessons (study 4)

This study builds on all the previous studies, in terms of theoretical tenets and methodology. Similar to study 3, I focus on the development of students' self-efficacy across a sequence of lessons in mathematics. On the basis of the factor structure for self-efficacy investigated through studies 1 – 3, I propose a new measure of students' mastery experiences, matched, item-by-item, with the measure for students' self-efficacy for classroom learning. In study 4 I use this instrument to investigate how students' mastery experiences are related to their self-efficacy, over the course of a series of lessons in mathematics. While the evidence for reciprocal effects between self-efficacy and mastery experiences is scarce overall, I know of no studies that investigated reciprocal effects between students' self-efficacy and mastery experiences during regular lessons in mathematics. There is a need to further our understanding of the process of self-efficacy change, involving cycles of mastery experiences and self-efficacy appraisals across a sequence of influential events. I conceptualise lessons in mathematics as potential self-efficacy changing events, and test the structural validity and invariance of my proposed measure of mastery experiences across 3-4 lessons. I investigate

stability and cross-lagged effects of self-efficacy and mastery experiences, and use equality constraints to compare the magnitudes of the effects. Finally, I test the validity of my proposed measures of self-efficacy and mastery experiences through investigating relationships with a previous established measure of ability conceptions. Similar to study 3, the data for study 4 stems from the phase two data collection.

1.2.5 Overall discussion and conclusion

In this chapter I discuss overall findings from my study. I discuss the implications from my research questions thoroughly in each chapter, thus I do not repeat these discussions, but rather focus on cross-study findings and implications, and directions for future research in light of these. I emphasise important overall limitations to my findings (limitations are also discussed in each study), and conclude with the contributions each of my studies have made to the overarching themes of my thesis.

1.3 Brief theoretical background

In this section I present the overall theoretical background to the thesis. In order to identify and build on literature relevant to this thesis, I started with central review papers related to research on self-efficacy and mathematics education (e.g. Hannula, 2012; Klassen & Usher, 2010; Pajares, 1996; Usher & Pajares, 2008). For each key paper identified from these reviews, I would follow up on references to previous papers, as well as use the "cited by" function in Google Scholar to identify more recent papers. While this approach to reviewing the literature led to many papers relevant to the themes of my thesis, I complemented this list with searches in several data bases (EBSCO, ERIC, Google Scholar, Mendeley, PsycINFO, Scopus, Web of Science), using relevant search terms (e.g. self-efficacy*, math*, student*, level*, change*). I repeated this process of literature search and review for each of the four studies in the thesis, adapting the search terms according to the themes of each study. Not

every study I identified and read is included in this thesis. However, I have attempted to include the most relevant literature for each of the four studies. The overview afforded by my broad approach to the literature review has aided this process of selection.

For each of the studies the thesis is made up of, I will further discuss theory relevant to the themes in each study, while in this section I will concentrate on the overall themes. I have organised this section according to previous research on: the conceptualisation of self-efficacy; the relationship between self-efficacy and performance; and the development of self-efficacy. On the basis of knowledge stemming from the theoretical and empirical studies I review, I will present a theoretical model for my thesis as a whole (section 1.4).

1.3.1 Conceptual underpinnings of self-efficacy

The construct of self-efficacy (see section 1.1 for a definition) was first introduced by Albert Bandura in his social cognitive theory (1977). Bandura proposed there is a triadic reciprocal relationship between individuals' cognitions, affect and biology (internal, personal factors), their behaviour, and their external environment (Bandura, 1997). This model of triadic reciprocal causation highlights that people are both producers and products of social systems, and that self-referent thoughts are important in regard to how people respond and adapt to environmental events.

Self-efficacy differs from other self- and expectancy beliefs such as self-esteem, self-concept and outcome expectancy in the degree of context specificity and task correspondence: it is sensitive to the shifting demands and environmental conditions that influence successful task performance under different circumstances (Pajares, 1996b; Valentine & DuBois, 2005). In contrast, self-concept and self-esteem involve "general appraisals based on accumulated

experiences across multiple sources" (Schunk & Miller, 2002, p. 32), and are thus related to more general appraisals of the self.

Self-efficacy is proposed to vary according to three dimensions: level, strength, and generality (Bandura, 1997; Pajares, 1996b; Zimmerman, 2000). Level refers to the level of perceived difficulty of tasks, strength refers to the degree of confidence about performing successfully, and generality refers to the degree of specificity (versus generality) of the expectations. In the current study, I refer to facet-specificity in place of generality, as this aligns with my focus on facets of mathematics.

Studies have investigated mathematics self-efficacy in regard to a variety of facets, such as problem solving (Pajares, 1996a; Schukajlow et al., 2012) and self-regulated learning (Pajares & Graham, 1999; Zimmerman & Kitsantas, 2005), and numerous studies have reported on the relationship between self-efficacy strength and performance (see for instance Multon et al., 1991). However, while Bandura advised self-efficacy measurements should include gradations of challenge: "Perceived efficacy should be measured against levels of task demands that represent gradations of challenges or impediments to successful performance." (Bandura, 2006, p. 311), we know little about students' mathematics self-efficacy in regard to tasks of varying perceived difficulty. Chen and Zimmerman (2007) used students' performance results to determine level of difficulty post-hoc, and found that students' mathematics self-efficacy was the least predictive of performance for difficult tasks, i.e. the correlation was lower for self-efficacy and performance on the tasks the students did badly at. In contrast, Locke et al. (1984) included level of difficulty in the self-efficacy items a priori, and found that the correlation with performance was highest for tasks of medium to hard perceived difficulty, when investigating memory self-efficacy. The different results from

these studies may be related to their differing conceptualisation of levels of self-efficacy. In my study I include level of difficulty in the self-efficacy items a priori, given I am interested in perceived rather than actual level of difficulty of the tasks.

Measurement issues have been common in self-efficacy studies (Pajares, 1996b). Klassen & Usher (2010) found that out of 109 articles reviewed from some of the most prolific journals (in regard to publishing articles about self-efficacy), more than 50% of the studies were deemed "not congruent" to the theoretical tenets and measurement recommendations from self-efficacy theory. Efficacy beliefs measurements in the articles deemed non-congruent often suffered from a lack of specificity, and some were lacking in resemblance to the theoretical concept of self-efficacy, assessing instead related concepts such as outcome expectancies or self-concept. It is difficult to assess the contribution to knowledge when it is unclear which constructs are being measured, which highlights the importance researchers are clear how they conceptualise the self-efficacy construct. In this study I use the most common conceptualisation of self-efficacy as proposed by researchers such as Bandura (1997), Zimmerman (2000), Pajares & Miller (1995), and Klassen & Usher (2010).

To sum up, while level of perceived difficulty is central to the commonly accepted conceptualisation of self-efficacy, it does not appear any instruments have been designed that include level of perceived difficulty when measuring students' self-efficacy in mathematics.

1.3.2 Relationship between self-efficacy and performance

The findings from several meta-analyses indicate there is a significant positive relationship between self-efficacy and performance. In their analyses of 39 studies Multon, Brown and Lent (1991) found overall positive and statistically significant relationships between self-efficacy and academic performance, with a mean effect size of $r = .38$. However, effect sizes

varied according to study characteristics. For instance, the positive relationship between self-efficacy and achievement outcome was weaker for elementary school students than high school or college students. Valentine, Dubois and Cooper (2004) performed a meta-analyses investigating findings from 55 studies on the relationship between different types of self-beliefs (self-esteem, self-concept and self-efficacy) and achievement. They included only studies controlling for prior achievement, and found an overall positive effect size of $r = .08$. The authors reported that effect sizes from studies on self-efficacy tended to be larger than from studies on other self-beliefs, but they could not test this relationship directly, as there were too few self-efficacy studies available. However, the authors found that differences in effect sizes were related to the level of specificity the measures were most likely to be assessed at (self-efficacy was more likely to be assessed at the academic than the global level). Studies that included global measures of self-beliefs wielded smaller effect sizes ($\beta = .07$) than studies including academic measures ($\beta = .13$). Furthermore, they found larger effect sizes associated with studies where self-belief measures were matched with achievement measures ($\beta = .11$) than when they were not matched ($\beta = .06$) (Valentine et al., 2004).

In regard to mathematics, self-efficacy is related to students' choice to engage with difficult tasks (Bandura & Schunk, 1981), persistence in problem solving (Schunk, 1981), anxiety (Pajares & Kranzler, 1995), and use of learning strategies (Zimmerman & Martinez-Pons, 1990). These motivations and behaviours are related to learning and performance in mathematics. Self-efficacy has been found to predict mathematical problem-solving after controlling for factors such as self-concept, perceived usefulness of mathematics, prior experience with mathematics, cognitive ability, mathematics Grade Point Average, anxiety, and gender (Pajares, 1996a; Pajares & Miller, 1994). Mathematics self-efficacy has been investigated in a range of countries, for instance by Schulz (2005), who reported on the PISA

2003 results. Schulz found that across the OECD (The Organisation for Economic Co-operation and Development) countries included in his sample, self-efficacy was positively correlated with mathematics literacy (OECD median= .52, Norway= .55), and negatively correlated with mathematics anxiety (OECD median= -.34, Norway= -.50).

Social-cognitive theory proposes a reciprocal relationship between self-efficacy and achievement experiences (Bandura, 1989). While numerous studies have investigated self-efficacy and achievement correlations, relatively few have investigated the reciprocal relationship between self-efficacy and mathematics achievement empirically. Reciprocal effects between self-beliefs and academic achievement have been proposed for several types of self-beliefs (Valentine & DuBois, 2005). In a recent meta-analysis Talsma, Schüz, Schwarzer, and Norris (2018) investigated reciprocal effects between academic self-efficacy and academic performance. The authors analysed the data from 11 studies and found support for a reciprocal effects model. While 6 of the 11 studies considered mathematics, none of the original studies had investigated reciprocal effects between mathematics self-efficacy and performance in the case of school age children. Thus only few studies have investigated reciprocal effects between self-efficacy and performance in mathematics, and seemingly none including data from Norway. Williams and Williams (2010) proposed a reciprocal effects model, which fit the data well for PISA results from 26 out of 33 countries. They argued reciprocal effects might be a fundamental psychological process, however the data from Norway was excluded from the analyses due to little variation in the grade-level variable.

The Norwegian school context differs in some important ways to that of e.g. the UK, in that more than 96% of students attend state schools (Statistics Norway, 2017), and permanent ability grouping is illegal (Opplæringslova, 1998, § 8-2). In other words, classrooms are

largely characterised by a mixture of ability grouping, as well as relatively mixed student backgrounds in socio-economic terms. Conducting investigations in this context might be advantageous in terms of generalisability of results, as there are few systematic differences between schools.

To sum up, empirical findings indicate there is significant relationship between self-efficacy and performance, and that the magnitude of the relationship varies according to study, student and task characteristics. While strong empirical evidence is found for a positive relationship between mathematics self-efficacy and students' mathematics achievement, the hypothesized reciprocal relationship between self-efficacy and mathematics has been tested in only few studies, and not with a Norwegian student sample. Furthermore, seemingly no studies have investigated reciprocal effects including level, strength, and facet-specific self-efficacy.

1.3.3 Development and change of self-efficacy

Bandura has proposed four factors as sources of self-efficacy (1997): mastery experiences, vicarious experiences, social persuasion and physiological state. While mastery experiences are posited as the most important source, providing the most resilient self-efficacy beliefs (Byars-Winston, Diestelmann, Savoy, & Hoyt, 2017; Jungert & Andersson, 2013), the relative importance of the other sources on self-efficacy is uncertain. The sources of self-efficacy do not influence self-efficacy directly, rather, individuals make cognitive appraisals or interpretations of the information gained through the sources, and evaluate what it means in regard to their ability to perform certain tasks (Stajkovic & Sommer, 2000; Usher & Pajares, 2008). Lindsley, Brass, and Thomas (1995) argued self-efficacy is cyclically related to performance, where performance influences self-efficacy, which influence subsequent performance (through students behaviours and affect), and so on. As such, self-efficacy develops on the basis of the most recent self-efficacy changing information (Shea & Howell,

2000), as an effect of *influential events* such as mastery experiences. In this view mastery experiences do not reflect students' actual or objective performances, but rather students' interpretation of their performance experiences and what these mean.

Once self-efficacy beliefs have been formed, their resilience depends on how they were acquired, how strong they are and the nature of any disconfirming experiences (Bandura, 1997; Schunk & Miller, 2002). In a study by Usher and Pajares (2006) mastery experiences did not predict the self-efficacy of low-achieving students. The authors argued that just like students who have built up strong self-efficacy due to many positive mastery experiences might be relatively unaffected by the occasional set-back, the self-efficacy of students who have experienced many failures might be less susceptible to change by an occasional mastery experience (Bandura, 1997). Similarly, Butz and Usher (2015, p. 59) argued that students who had come to see themselves as less academically capable were "predisposed to look for confirming evidence of their own inefficacy".

Student age, task difficulty, and conceptions of ability are some factors likely to interact with the relationship between sources of self-efficacy and self-efficacy, through either affecting the types of experiences individuals have, or affecting their cognitive appraisal of these experiences. Differences in self-efficacy have been found for students of different ages (Bong, 2009; Davis-Kean et al., 2008). Younger students are less apt at assessing feedback information and task requirements than older students. Nicholls (1978) described the cognitive development of pre-adolescents. He argued that changes in regard to children's developing perceptions of effort, ability, and level of task difficulty were related to age increases in stability and accuracy of self-perceptions. Several researchers have found the self-beliefs of younger students to be more uniform (failing to differentiate between different

aspects of the self), and that they tend to display overly optimistic and less "accurate" self-efficacy than older students (Bong, 2001, 2009; Davis-Kean et al., 2008; Multon et al., 1991; Schunk & Pajares, 2010). Self-beliefs tend to become more differentiated, negative and more closely associated with achievement by age (Valentine et al., 2004; Wigfield et al., 1997).

The tasks students engage with might serve to influence students' self-efficacy through their previous task experiences. Similarly, students' perceptions of future tasks can influence the extent to which they believe they can successfully accomplish said tasks. Self-efficacy might be influenced by students' perceptions of task difficulty (Lodewyk & Winne, 2005), and also how much help they received while working on the task. In a meta-analysis Stajkovic and Luthans (1998) demonstrated that the level of task complexity (objectively determined) moderated the relationship between self-efficacy and work-related performance, where the strength of the relationship decreased with increasing complexity. In my investigations I am primarily interested in students' subjective perceptions of difficulty, which may differ from task difficulty as objectively determined (Dowling, 1978).

Ability conceptions are closely aligned with central tenets of social cognitive theory. Students may perceive of ability to be relatively more stable and unchanging (entity beliefs), or relatively more dynamic and related to effort (incremental beliefs) (Dweck, 2008; Dweck & Leggett, 1988). There is a relationship between conceptions of ability and self-efficacy, where ability conceptions might influence students' cognitive appraisal of efficacy-building information (Schunk & Zimmerman, 2007), as well as how they cope in performance situations (Davis, Burnette, Allison, & Stone, 2011). Butler found that individuals' beliefs about ability moderated their perceptions of feedback, and whether they placed more emphasis on temporal or normative feedback information (Butler, 2000). Conceptions of

ability are also related to student age, where younger students have been found to report higher beliefs in effort (as opposed to ability) than older students (Schunk & Lilly, 1984), mirroring their developing self-perceptions.

A limited number of studies have investigated the development of students' self-efficacy over time. Caprara and colleagues investigated development of self-efficacy for self-regulated learning, and reported a progressive decline in students' (12-22 year olds) self-efficacy over time (Caprara et al., 2008). Changes in students' self-efficacy during periods of transition in school have previously been investigated by Midgley and colleagues in several studies (Friedel, Cortina, Turner, & Midgley, 2010; Midgley, Feldlaufer, & Eccles, 1989). They demonstrated teacher and classroom characteristics influenced changes in students' self-efficacy. Schunk and Pajares (2010) argued students' self-efficacy is likely to be affected by periods of transition, and typically start to decline during middle school, however only few studies I am aware of have investigated students' self-efficacy with longitudinal data, and the findings are not conclusive (e.g. Caprara, Vecchione, Alessandri, Gerbino, & Barbaranelli, 2011; Hornstra, van der Veen, Peetsma, & Volman, 2013).

There may be different trends of self-efficacy change across different time frames, and within different learning contexts. Phan (2012) investigated developmental trajectories of mathematics self-efficacy for 3rd and 4th grade students in Australia. During four occasions of measurement over 1 year, students' self-efficacy increased (Phan, 2012). Even more fine-grained, Bernacki et al. (2015) investigated learners' self-efficacy and problem solving performance during a learning task consisting of several problems in an intelligent tutoring system. Within a single learning task, they found variations in self-efficacy for 60% of the students. Furthermore, increases in self-efficacy predicted help seeking, performance and

learning when working on the next problem. The approach by Bernacki et al. (2015) was fully automated, where students interacted with a cognitive tutor (rather than with a teacher during regular lessons). More research is needed to fully understand the development of students' self-efficacy at different ages and during different time-spans, and during regular lessons in mathematics.

This section has demonstrated there is theoretical and empirical support that mastery experiences are the most influential source of students' self-efficacy. Students evaluate information stemming from the sources of self-efficacy cognitively, where the relationship between mastery experiences and self-efficacy takes a cyclic form of reciprocal effects. Furthermore, changes in self-efficacy is likely moderated by factors such as student age, task difficulty, and ability conceptions. Some studies have indicated students' self-efficacy beliefs decrease progressively over time, when comparing students in different age groups. There is little research into the developmental trajectories of students' self-efficacy on a smaller time-scale, where one study has indicated students' self-efficacy varies during learning tasks. We do not know how students' self-efficacy changes from day to day during regular lessons in mathematics.

1.4 A model of students' self-efficacy in mathematics

On the basis of existing knowledge about self-efficacy a theoretical model is proposed (Figure 1). This model presents a theoretical view of the relationship between level, strength, and facet-specific self-efficacy in mathematics, performance in mathematics, and students' experiences on mathematical tasks. In reference to the triadic model of reciprocal causation presented by Bandura (1997, p. 5), my model focuses on the individual factors related to students' perceptions and experiences, while it is understood that these experiences have a reciprocal relationship also with external factors such as families, teachers and classrooms,

not investigated in the current thesis. Certain aspects of the model presented are strongly supported by the literature and by empirical findings. Other aspects of the model are theoretically proposed or have received some empirical support, but require further investigation. In the following paragraphs I summarise the theoretical model, drawing on the conclusions stemming from my discussion in the previous section. I indicate which aspects of the model are theoretically and empirically supported, and which are proposed, that are in need of further investigation.

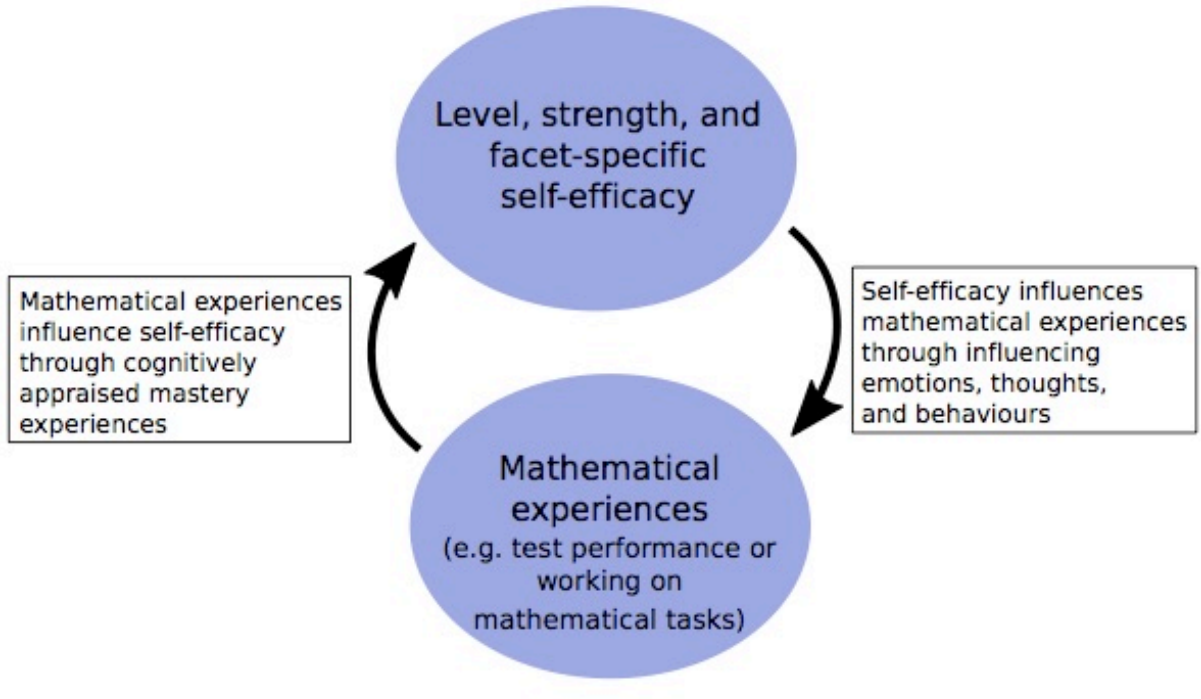


Figure 1: Theoretical model for the thesis

In section 1.3.2 I explained that self-efficacy influences students' emotions, thoughts and behaviours. Specifically, I discussed that mathematics self-efficacy is related to important learning behaviours such as engagement, persistence, and use of learning strategies, as well as their mathematics-related anxiety. These motivations and behaviours are subsequently related to performance in mathematics.

Furthermore, in section 1.3.1 I referred to several researchers who suggest self-efficacy is a multidimensional construct, and vary according to level, strength, and facet-specificity. Research has demonstrated that facet-specific self-efficacy (for instance problem solving self-efficacy, or self-efficacy for self-regulation) predict task performance, and that there is a correlation between self-efficacy strength and performance outcomes. I propose that self-efficacy also varies according to perceived level of difficulty, and that there is a relationship between level of self-efficacy and performance outcomes.

As I explained in section 1.3.2, there is a positive relationship between self-efficacy and performance. While several studies have demonstrated an effect of self-efficacy on performance in mathematics after controlling for ability or prior performance, the proposed reciprocal relationship between self-efficacy and performance outcomes in mathematics has been empirically tested in few studies only. I propose there is a reciprocal relationship between level, strength, and facet-specific self-efficacy, and national tests in mathematics.

In section 1.3.3 I explained that mastery experiences are the strongest influence on self-efficacy. Furthermore, I discussed there is a theoretically proposed reciprocal relationship between self-efficacy and mastery experiences, which is posited as a key factor in the process of self-efficacy change. I propose there is a reciprocal relationship between students' level, strength and facet- specific self-efficacy and mastery experiences, and that we can predict gains in students' self-efficacy through measures of their classroom mastery experiences.

To sum up, I propose there is a reciprocal relationship between level, strength, and facet-specific self-efficacy and mathematical experiences, where previous experiences (for instance

national test performance or classroom learning experiences) influence self-efficacy in the form of appraised mastery experiences, while self-efficacy influences learning and performance (on tests or mathematical tasks) through influencing students' emotions, thoughts and behaviours.

1.5 Contribution to knowledge

I will now discuss how the present study will contribute to knowledge, by addressing aspects of the theoretical model (section 1.4) that are in need of further investigation. Finally, I will state the research questions pertaining to this thesis. The overall contribution of this thesis is twofold, with one methodological and one substantive strand. First, I propose a new measure of self-efficacy, taking into account the multidimensional conceptualisation of the construct. Second, I propose new methodologies to measure the development of students' self-efficacy in the mathematics classroom, using mirrored instruments of self-efficacy and mastery experiences.

These new proposed methodologies allow for new avenues of investigations, to contribute to knowledge in substantive terms. Specifically, I will use the new multidimensional measure of self-efficacy to investigate the relationship between level, strength, and facet-specific self-efficacy and students' performance on national tests, both in a cross-sectional and a longitudinal study. Further, I will use repeated applications of a new measure of mastery experiences, which is matched item-by-item with a measure of self-efficacy, to investigate the relationship between mastery experiences in mathematics and the development of students' self-efficacy over a series of lessons in mathematics. I will now expand on the different ways my thesis will contribute to knowledge, considering current knowledge in the field (which was discussed in the brief theoretical background, see section 1.3).

Students' self-efficacy is related to the tasks or challenges considered when they formulate their beliefs. Students' self-efficacy, and the relationship between their self-efficacy and performance might differ as a function of perceived level of difficulty of tasks. Theorists such as Bandura have suggested measures of self-efficacy should include gradations of challenge, however it does not appear as if an instrument measuring self-efficacy in mathematics including levels of perceived difficulty has been tested empirically (see section 1.3.1). The present thesis will contribute to knowledge through assessing the structural validity of a new instrument that includes level, strength, and facet-specific self-efficacy, and their relationship with performance outcome.

The empirical link between self-efficacy and academic outcome has been replicated in many studies, and further studies have supported a directional effect of self-efficacy on achievement in mathematics also when controlling for prior achievement (see section 1.3.2). While a reciprocal effects model has received ample attention and empirical support within self-concept research, the theoretically proposed reciprocal relationship between students' mathematics self-efficacy and academic achievement has received little attention empirically. Furthermore, the reciprocal effects model has seemingly not been investigated when taking into account level as well as strength and facet-specific self-efficacy. The current thesis will contribute to knowledge through investigating empirically the reciprocal relationship between level, strength, and facet-specific self-efficacy and students' performance outcomes on national tests in mathematics in a longitudinal design.

Some research studies have found that students' self-efficacy declines with age, while simultaneously increasing in predictive power in relation to achievement outcomes (see section 1.3.3). At the same time studies have reported that students' self-efficacy increases

over shorter time spans, and the development of self-efficacy in mathematics for students of different ages is uncertain. I have located only one study that investigated changes in students' self-efficacy through frequent assessments over a shorter time span (Bernacki et al., 2015). The study measured students' self-efficacy and performance while using an intelligent tutoring system. The current thesis will contribute to knowledge through investigating systematically the stability and change of students' self-efficacy across a sequence of regular classroom lessons when students are introduced to new topics.

Mastery experiences are viewed as the strongest source of self-efficacy, where the reciprocal relationship between self-efficacy and mastery experiences is theoretically proposed as key to the process of self-efficacy change (see section 1.3.2). At the same time, I have found no studies that investigated the reciprocal relationship between students' mathematics self-efficacy and mastery experiences across learning events. My thesis will contribute to knowledge through investigating empirically the relationship between students' self-efficacy and mastery experiences across a sequence of lessons in mathematics, while students learn new mathematical topics.

1.5.1 Research questions

I started this thesis by presenting two overarching questions: What is the relationship between the kinds of experiences students have with mathematics, and their confidence that they can perform well on mathematical tasks? and: How does students' confidence for future learning and performance in mathematics develop during classroom lessons? Following on from my overarching questions and the proposed theoretical model of self-efficacy in mathematics, I present 11 research questions, which I will address in four studies. Each study is stand-alone, yet related through the overarching questions. Furthermore, each study builds on the investigations from the previous studies.

Study 1:

(1) *What is the structural validity of the proposed self-efficacy constructs, which include facets and levels of self-efficacy?*
(2) *How is facet-specific and level-specific self-efficacy associated with performance on national tests in mathematics?*
(3) *What is the relationship between students' self-efficacy for perceived easy, medium, and hard tasks, and the students' scores on easy, medium, and hard level items on national tests in mathematics?*

Study 2:

(1) *What is the relationship between self-efficacy and performance on national tests in mathematics in grade 8, and self-efficacy and performance on national tests in mathematics in grade 9?*
(2) *Do the relationships between self-efficacy and performance on national tests differ according to levels or facets of self-efficacy?*

Study 3:

(1) *What is the structural validity of the proposed self-efficacy constructs, which include facets and levels of self-efficacy?*
(2) *How stable is the rank-order of students' mathematics self-efficacy across a sequence of classroom lessons in mathematics?*
(3) *How does students' mathematics self-efficacy change across a sequence of classroom lessons in mathematics?*

Study 4:

(1) *What is the structural validity of the proposed mastery experiences constructs, which include facets and levels of mastery experiences?*
(2) *What is the relationship between self-efficacy and mastery experiences across a sequence of regular lessons in mathematics?*
(3) *What is the relationship between students' self-efficacy, mastery experiences, and ability conceptions?*

2 Level, strength, and facet-specific self-efficacy, and students' performance on national tests in mathematics (study 1)

Self-efficacy is an important motivational construct as it is associated with student engagement and learning experiences in different subject areas, including mathematics. Self-efficacy beliefs have received extensive attention since they were introduced in the 1960s by Albert Bandura, and numerous studies have measured students' mathematics self-efficacy. Going beyond previous operationalisations of self-efficacy I propose a new measure that takes into account not only facet-specificity (beliefs related to particular competences or skills) and strength (confidence of the beliefs), but also level (perceived task difficulty) of these beliefs as proposed by Bandura (1997, 2006). In particular, I included level-specific items referring to perceived difficulty on a subsequent national test in mathematics. In total 756 Norwegian grade 5, 8, and 9 students completed the "Self-Efficacy Gradations of Difficulty Questionnaire" that I developed. I tested plausible multitrait-multimethod models using structural equation models. The best fitting model included three factors representing levels of perceived difficulty, and a-priori specified correlated uniquenesses representing the following four facets that related to problem solving or students' self-regulation skills during the test: (1) to complete a certain number of problems, (2) to solve tasks of a certain challenge, (3) to concentrate, and (4) to not give up for a certain amount of time. The results indicated that both test facets and levels of perceived difficulty are related to students' self-efficacy, which in turn is associated with scores on national tests in mathematics. Furthermore, students' self-efficacy for easy, medium, and hard tasks predicted their scores on easy, medium, and hard items of the national test.

The main contribution of this study is the presentation of a measure of mathematics self-efficacy that distinguishes between different levels of difficulty (as perceived by students), while previous measures have only considered strength and facets-specificity. Taking level (difficulty) into account broadens our understanding of the self-efficacy construct, because it allows the distinction between self-efficacy for different levels of perceived difficulty, as well as investigation into the relationship between self-efficacy and performance outcomes for each level. The study suggests a methodology to disentangle facet and level –specific self-efficacy, providing us with a tool for identifying students who differ in their self-efficacy for different facets or levels of difficulty of mathematics. Increasing our knowledge about mathematics self-efficacy for tasks of different types and levels of difficulty has implications in regard to the tasks educators are encouraged to provide for their students, and the kinds of support they offer during task engagement and learning.

2.1 Introduction

Self-efficacy is related to central learning behaviours and outcomes such as individuals' motivation and engagement (Martin et al., 2012), goal setting behaviours (Bong, 2009), persistence (Zimmerman & Ringle, 1981), and physiological experiences (Meece, Wigfield,

& Eccles, 1990). Self-efficacy has been investigated across a range of fields and domains, such as business, physical activity, neuroscience, medicine and education (Cairney, Hay, Faught, Leger, & Mathers, 2008; Dierdorff, Surface, & Brown, 2010; Gist, Schwoerer, & Rosen, 1989; Hirai et al., 2002; Klassen & Chiu, 2010; Lacker, Carosella, & Feuerstein, 1996; McConville & Lane, 2006; Miliaressis & Rompre, 1987; Rogers et al., 2006; Winfrey & Weeks, 1993). In mathematics in particular, self-efficacy has been linked with perseverance (Zeldin & Pajares, 2000), performance outcomes (Schulz, 2005), attitudes towards mathematics and choosing maths-related courses and careers (Hackett & Betz, 1989).

Self-efficacy differs when people are faced with different types of tasks or situations, and individuals consider a range of available information when they form their beliefs, such as their own previous performance on similar tasks, access to help, or the difficulty of the task. Accordingly, self-efficacy is presumed to vary according to both specificity and level of difficulty of the task. Context is thus related to the subject-specific task or facet in question, but it also includes situational demands that influence the perceived level of difficulty of the task. Measurement approaches that take account of contextual factors can broaden our understanding of these beliefs in important ways.

Previous studies have used a wide range of instruments to measure self-efficacy. Bong & Hocevar (2002) argued it might be difficult to compare findings from studies on self-efficacy given the range of methods applied to measuring them, however different measures have contributed in different ways to our understanding of self-efficacy and its relationship to other measures including performance scores. Existing instruments have incorporated *facet-specificity* (beliefs related to particular competences or skills) and *strength* (confidence of the

beliefs), while *levels* (perceived task difficulty) of mathematics self-efficacy have seemingly not been included explicitly as a research variable.

In this chapter, I contribute to existing research by proposing a new instrument of self-efficacy that takes into account not only strength and facet-specificity, but also level of self-efficacy as proposed by Bandura (1997, 2006). Taking level into account broadens our understanding of the self-efficacy construct, because it allows the distinction between self-efficacy for different levels of perceived difficulty, as well as investigation into the relationship between self-efficacy and performance outcomes for each level. In more detail, I will test the structural validity of my proposed multidimensional instrument, and I will investigate how level and facet-specificity of self-efficacy is related to subsequent performance on national tests in mathematics. To this end, I used a sample of 756 Norwegian grade 5, 8, and 9 students who completed the "Self-Efficacy Gradations of Difficulty Questionnaire" that I developed.

In the following sections, I first review Bandura's original proposals about the meaning of self-efficacy. I then outline facet-specificity, strength, and level of self-efficacy and discuss how these have been operationalised and measured in previous studies. After that I present my model that integrates facet-specificity, strength, and level of mathematics self-efficacy.

2.2 Theoretical background

Self-efficacy is individually held beliefs about personal capability to perform particular future tasks (Bandura, 1997). Albert Bandura first introduced the construct of self-efficacy in the 1960's. Bandura also provided the most widely used definition in the literature, that self-efficacy is individual beliefs people form regarding their own capabilities to perform particular tasks: "beliefs in one's capabilities to organize and execute the courses of action

required to produce given attainments" (Bandura, 1997, p. 3). While self-efficacy has become widely researched and many researchers have contributed to extending and clarifying our current understanding of the concept, researchers of self-efficacy such as Schunk (1989, 1991), Pajares (1996b), Zimmerman (2000) and Bong & Skaalvik (2003) refer back to versions of the original conceptualisation proposed by Bandura.

In the literature there is an array of theories concerning self-beliefs (Bong & Skaalvik, 2003; Pajares, 1992). Self-efficacy differs from other self-beliefs in its relation to specific, future challenges or tasks and its relative context specificity (Bong & Skaalvik, 2003; Pajares, 1996b; Valentine & DuBois, 2005), while constructs such as *self-concept* and *self-esteem* involve "general appraisals based on accumulated experiences across multiple sources" (Schunk & Miller, 2002, p. 32). These latter constructs are thus related to more general appraisals of the self, and/or the self-worth people associate with such accomplishments. Theorists further emphasise a mastery criterion of performance in relation to self-efficacy (example C*an I do this task?*), which stands in contrast to the normative comparisons proposed as an important source of self-concept (example *Am I better or worse than others at this task?*) (Bong & Skaalvik, 2003; Zimmerman, 2000). Self-efficacy items are commonly phrased in terms of *can* rather than *will*, as self-efficacy is conceptualised as related to perceptions of capability rather than intentions to act (Bandura, 1997, p.43).

Self-efficacy is also conceptually different to *outcome expectancies* (Bandura, 1997; Zimmerman, 2000) and *expectancy-value* theories (Schunk, 1991). Self-efficacy is related to people's confidence in their own capabilities of carrying out actions or producing certain accomplishments, which differs to the outcomes they expect from achieving such attainments (outcome expectancies) and the extent to which they value such expected outcomes (values).

I have chosen the construct of self-efficacy as it is relatively task-specific, and thus well suited to investigating the relationship between students' task-specific beliefs and their performance outcomes. Also, it may be useful to focus more on mastery and goal-referenced information (sources of self-efficacy) and less on normative comparisons (source of self-concept), to support students' performance (Bong & Skaalvik, 2003). Furthermore, previous studies have demonstrated that self-efficacy has equally high or higher correlations with performance outcomes, as compared with other self-beliefs (e.g., Bong & Clarke, 1999; Valentine, DuBois, & Cooper, 2004).

Self-efficacy is held to vary according to three dimensions: generality, strength, and level (Bandura, 1997; Pajares, 1996b; Zimmerman, 2000). *Generality* refers to the degree of specificity. For instance, self-efficacy for mathematics is more general as compared to self-efficacy for specific facets of doing mathematics, such as solving problems, but mathematics-self-efficacy is more specific than academic- self-efficacy. In this chapter, I use the term *facet-specificity* in place of "generality" given my theoretical model includes specific facets of taking a mathematics test. *Strength* refers to the amount of certainty about performance on a task (such as doing well on a national test in mathematics) or carrying out specific facets of test taking (such as persevering). *Level* refers to the degree of performance demand, difficulty, or challenge included in the task. For instance, it can refer to self-efficacy for tasks of increasing difficulty or conducted under increasingly difficult circumstances.

A variety of measures have been applied in studies investigating mathematics self-efficacy. Not all theorists ascribe to the conceptualisation of self-efficacy proposed by Bandura, and theorists who express similar conceptualisations of self-efficacy might differ in their

approaches to measurement. Existing measurement approaches vary mainly in regard to the degree of generality versus specificity of the construct, and in regard to the scale used to measure strength of self-efficacy at the designated level of specificity, while I have found only few measures of mathematics self-efficacy that include level of task demand. Next I elaborate on each of the three dimensions of specificity, strength, and level. I also include examples of previous measures that included each of the three dimensions. More space is afforded to the discussion about facet-specificity, as more studies have included this dimension, as compared with self-efficacy strength and level.

2.2.1 Facet-specificity

According to Bandura (1997) as well as other researchers (Bong, 2001; Pajares & Miller, 1995; Schukajlow et al., 2012), self-efficacy is a relatively specific construct, where individuals form their judgment based on an assessment of prospective future tasks and situations. As Bong puts it "the core in self-efficacy assessment, then, is to get to the very belief that individuals draw on when they are faced with situations that involve certain actions or performance" (2006, p. 293). According to this conceptualisation, global constructs such as general self-efficacy, referring to individuals' overall confidence across a range of domains, situations, and tasks (e.g. Scholz, Dona, Sud, & Schwarzer, 2002) are unlikely to address participants' subject-specific (e.g. mathematics-specific) self-efficacy, due to a lack of specificity (Pajares & Miller, 1995). Despite the argued context specificity of self-efficacy, some researchers have highlighted the usefulness of more general measures of self-efficacy. Scholz and colleagues (Scholz et al., 2002) tested the psychometric properties of the General Self-Efficacy scale (example item *"Thanks to my resourcefulness, I can handle unforeseen situations"*), which was related to their conceptualisation of self-efficacy as "global confidence" at its most general level.

One example of a general measure of self-efficacy used in relation to a study of mathematics is the Student Motivation Scale. The Student Motivation Scale is a large instrument developed by Martin (2003) which consists of 6 "boosters" (hypothesized to increase motivation) and 4 "guzzlers" (hypothesized to reduce motivation) as subscales. Martin (2003) defined self-efficacy as "students' belief and confidence in their ability to understand or to do well in their schoolwork, to meet challenges they face, and to perform to the best of their ability", which is similar to the definition of general self-efficacy given by Scholz and colleagues (2002), although specified for an academic context. Martin refers to self-efficacy theory and Bandura (1997) to include "self-beliefs" as a booster (example item *If I try hard, I believe I can do my schoolwork well*). Each booster and guzzler consists of four items, which are rated on a 7-point Likert-type scale from 1 (Strongly disagree) to 7 (Strongly agree). The self-efficacy measure is thus of a general (academic) type, while the performance measures in the 2003 study consisted of school test scores in English and Mathematics (for students in years 9 – 11) (Martin, 2003). The measure of self-efficacy applied in this instrument seems to reflect the author's desire to present a unified instrument including different theoretical perspectives of motivation. After aggregating participants' scores for each of the ten subscales, Martin reported significant but modest effects between self-beliefs and maths achievement (Pearson's r =.20) and numeracy (Pearson's r =.15).

Bong (2001) argued that general self-efficacy might not function in the same manner as subject-specific beliefs, and that subject-specific self-efficacy cannot be inferred from higher-order (more generalized) self-efficacy. Bong (2001) tested the hierarchical representation and between-subject relations of self-efficacy in four school-subjects (Korean, English, mathematics and science). The results indicated a second-order hierarchical structure of students' self-efficacy in the four subjects, and Bong (2001) argued self-efficacy is subject-

specific, yet moderately correlated across subjects. Older students' self-efficacy beliefs were more subject-specific than those of younger students, demonstrated in lower cross-subject correlations of self-efficacy (median value .55 for middle school sample, .42 for high school sample). Bong provided further support for the subject-specificity of self-efficacy in a subsequent study (2002), and furthermore reported that subject-specific, task-specific and problem-specific self-efficacy formed correlated, yet independent factors within each of the academic domains English and mathematics. Bong (2002) argued this supported Bandura's (1997) theory that self-efficacy can be distinguished according to three main levels of specificity, where the most general level is self-efficacy at the domain (subject) level when there is no reference to specific activities or situational demands.

Several examples exist in the literature of measures for mathematics self-efficacy at the subject level of specificity, for instance The Motivated Strategies for Learning Questionnaire (MSLQ) developed by Pintrich and De Groot (1990), and The Patterns of Adaptive Learning Scale (PALS) developed by Midgley et al. (2000).

The Motivated Strategies for Learning Questionnaire (Pintrich and De Groot, 1990) contains a self-efficacy subscale and has been widely used across different academic subjects (Bråten et al., 2004; Kıran & Sungur, 2011; Lodewyk & Winne, 2005; Muis & Duffy, 2013; Phan, 2013; Pintrich, Smith, Garcia, & McKeachie, 1993) including in relation to mathematics self-efficacy (Bong, 2009; Schweinle & Mims, 2009; Yetkin Özdemir & Pape, 2013). The MSLQ includes items related to motivation, cognitive and metacognitive strategy use and management of effort. Pintrich and De Groot (1990) referred to a general expectancy-value model of motivation as theoretical framework for the MSLQ, and used self-efficacy to address the expectancy component of the model. These beliefs were conceptualised as

"students' beliefs that they are able to perform the task and that they are responsible for their own performance" (example item *I expect to do very well in this class*). The self-efficacy subscale in the original MSLQ measure consists of nine items related to students' perceptions of competence and confidence in class work, rated on a 7-point Likert-type scale (1=not at all true of me, 7=very true of me). Four of these items are normative comparisons (example item *My study skills are excellent compared with others in this class*).

While the original authors of the MSLQ made no explicit reference to Bandura's conceptualisation of self-efficacy, several researchers using the MSLQ have referred to Bandura in their studies. Bong (2009, p. 884) referred to self-efficacy as "students' subjective convictions for successfully carrying out tasks in math class" (example item *I am sure that I can do an excellent job on the problems and tasks assigned for math class*), and reported that the correlation between self-efficacy and performance measures increased with student age from lower elementary to middle school. Bong had excluded 3 (normative) items from the MSLQ for "theoretical reasons".

The Patterns of Adaptive Learning Scale (Midgley et al., 2000) is another large instrument with a subscale measuring academic self-efficacy. The PALS student scale includes 94 items, organized in five different sections (Personal achievement goal orientations, Perception of teacher's goals, Perception of classroom goal structures, Academic-related perceptions, beliefs and strategies, and Perceptions of parents, home-life and neighbourhood). The Academic Efficacy subscale includes 5 items related to "students' perceptions of their competence to do their class work" (example item *I'm certain I can master the skills taught in class this year*) (Midgley et al., 2000, p. 20). Midgley et al. (1998) conceptualised self-efficacy as "students' confidence in their ability to master new skills and tasks" referring to

Pajares & Miller (1994), who in turn based their conceptualisation of self-efficacy on Bandura (1986). Several researchers have adapted the academic efficacy subscale of PALS to a math context (e.g. Fast et al., 2010; Friedel et al., 2010; Hornstra et al., 2013; Rimm-Kaufman, Baroody, Larsen, Curby, & Abry, 2015). Rimm-Kaufman and colleagues adapted the items to a mathematics classroom setting (example item *I'm certain I can master the skills taught in math this year*) defining math self-efficacy as "students' perception of their capacity to learn or perform in math" (Rimm-Kaufman et al., 2015, p. 173). The authors argued students' self-efficacy was significantly related to their engagement in math class through student self-report and as reported by the teacher (Rimm-Kaufman et al., 2015).

One benefit of conceptualising self-efficacy in subject related (as opposed to task-specific) terms, is that the same measures might be used across different studies and/or subjects. Bonne (2012) argued this would eliminate the resource-demanding process of piloting, while other benefits could be cross-study and cross-subject comparability of findings. According to Bandura (1997) domain-level self-efficacy is a good predictor of performance, however domain-level "indefinite items still sacrifice explanatory and predictive power" (p. 48).

Self-efficacy might also vary across different facets within an activity domain (Bandura 2006), and Pajares (1996b) argued that even subject-specific measures can fail to prompt participants to have a specific task or activity in mind when responding. Within the domain of mathematics, self-efficacy may be specified in regard to learning mathematics, sub-areas of mathematics, mathematical competencies, or specific mathematical tasks (Schukajlow et al., 2012).

Bandura (2006, pp. 324, 326) proposed instruments to measure facet-specificity of self-efficacy, such as problem-solving self-efficacy and self-efficacy for self-regulated learning. To gauge participants' problem solving-specific self-efficacy they were asked to rate how certain they were that they could solve a certain proportion of academic problems, e.g., 10% of the problems, 20% of the problems, and so on up to 100% of the problems. Participants' self-efficacy for self-regulation was gauged by asking how certain they were that they could concentrate on school subjects during class, and persevere in their schoolwork.

Lodewyk and Winne (2005) adapted items from the *expectancy* component of the Motivated Strategies for Learning Questionnaire developed by Pintrich and De Groot (1990), to measure self-efficacy for learning (example item: *I'm confident I am understanding the most difficult material in this project)* and self-efficacy for performance (example item: *I expect to do well on this project*) in their study about learning tasks in the classroom. In this case, self-efficacy for learning, but not for performance, strongly predicted students' performance on a well-structured science task, after controlling for prior general achievement. The aforementioned questionnaire has been used widely to investigate self-efficacy across different academic domains (e.g. Muis & Duffy, 2013; Phan, 2013), including mathematics (Schweinle & Mims, 2009; Yetkin Özdemir & Pape, 2013). The items used by Lodewyk and Winne (2005) to assess self-efficacy for performance are related to expectancies, which is not congruent with the main theoretical understanding of self-efficacy. The latter may explain the different findings for self-efficacy for learning and self-efficacy for performance in Lodewyk and Winne's study.

According to Pajares and Miller (1995), the specificity of the measure for self-efficacy should be aligned with the research question, as well as any performance measures used. A number

of studies have tailored self-efficacy measures to performance tasks through letting participants briefly see examples of the problems that are to be solved, prior to rating their confidence in solving them (Bandura & Schunk, 1981; Betz & Hackett, 1983; Bonne, 2012; Norwich, 1987; Pajares & Miller, 1995; Pampaka et al., 2011; Shih & Alexander, 2000; Yailagh, Lloyd, & Walsh, 2009) Schukajlow et al. 2012). This strategy enables high correspondence between self-efficacy and the performance measures. Pajares (1996a) found that task-specific mathematics self-efficacy predicted performance on those particular tasks for gifted ($\beta = .46$) and regular education ($\beta = .39$) students. For both samples, self-efficacy for self-regulated learning predicted mathematics self-efficacy ($\beta = .42$ and $\beta = .37$, respectively), but not performance. Mathematics self-efficacy was measured by first displaying the tasks to the participants, before asking them to rate their confidence in solving them. Self-efficacy for self-regulated learning was measured with items from the Children's Self-Efficacy for Self-Regulated Learning scale (Bandura, 2006), which is related to general schoolwork (e.g. organise and plan schoolwork, and concentrate during class). In the study by Pajares (1996b) the measure for problem solving self-efficacy was more specific than the measure for self-efficacy for self-regulation, while in my study items related to problem solving self-efficacy and self-efficacy for self-regulation are equally specific.

The first instrument which included confidence items matched with following performance assessments was the Mathematics Confidence Scale (MCS) created by Dowling (1978). This measure includes 18 items related to solving arithmetic, algebra and geometry problems at different levels of difficulty (computation, comprehension and application), across two different contexts (real and abstract) (example item related to 'abstract application of algebra': *There are three numbers. The second is twice the first and the first is one-third of the other number. Their sum is 48. Find the largest number.*). While the problems were supposed to

represent three hierarchical levels of cognitive demand, it is possible to construct problems within each of these three categories at a range of levels of difficulty. In fact, Dowling reported that not all students perceived the differences between the levels as intended. In the original scale, participants were asked to report their degree of confidence in solving each case using the letters A (no confidence at all) to E (complete confidence).

In original or adapted form, the MCS has been widely used over many years (Pajares & Miller, 1994). The MCS was included as a subscale in the Mathematics Self-Efficacy Scale (MSES), developed by Betz and Hackett (1983), a measure often employed in self-efficacy studies (i.e. Hackett & Betz, 1989; Lent et al., 1993; Randhawa, Beamer, & Lundberg, 1993). Betz and Hackett operationalized mathematics self-efficacy in a wide sense, including confidence to solve mathematical problems, perform everyday mathematical tasks, and do well on mathematics-related courses. At the same time the authors referred to self-efficacy theory as postulated by Bandura (Bandura, 1986), and argued that self-efficacy should be measured in relation to specific tasks (Hackett & Betz, 1989, p. 264). The MSES includes 52 items organized across the Mathematical Task subscale (example item *Balance your checkbook without a mistake*), the Mathematics Courses subscale (example item *Advanced calculus*) and the Mathematics Problems subscale (example item *The average of three numbers is 30. The fourth number is at least 10. What is the smallest average of the four numbers?*), with the latter subscale being adapted from the MCS. For items in each subscale participants are asked to indicate their confidence to perform the task, achieve grade "B" or better, or to solve the problem, respectively. In the original use of the MSES participants were asked to rate their confidence on a 10-point scale from 0 (No confidence) to 9 (Complete confidence), and scores were calculated for each of the subscales, as well as overall (Betz & Hackett, 1983; Hackett & Betz, 1989). Hackett and Betz (1989) used 18 items matched to the

MCS subscale as a measure of performance and the overall MSES score as a measure of self-efficacy, and reported a correlation of .44 between these two measures. In comparison, Pajares and Miller (Pajares & Miller, 1994) found a correlation of .70 between the MCS and a performance measure corresponding with the items in the MCS.

The MSES has since been altered by Pajares and Miller (1995) to form the Mathematics Self-Efficacy Scale Revised (MSES-R), which has also been widely used (Ayotola & Adedeji, 2009; Peters, 2013). The MSES-R differs from the original MSES in that it includes slightly different items in the Mathematics Problems Subscale, reflecting items included in the original, final version by Dowling (1978). Also, participants are asked to rate their confidence on a 5-point Likert-type scale.

An alternative to presenting representations of actual problems is to provide verbal or written descriptions of problems/tasks, and ask participants to rate their confidence to solve each one. Bong (2002) referred to task-specific self-efficacy as "verbal descriptions of tasks (…) written to reflect essential skills or concepts generally required across problems of similar types" (p. 138). The Mathematical Task subscale of the MSES mentioned above would be one example of this. This approach is similar to presenting specific problems in that specific "anchors" are provided to estimate self-efficacy. Simultaneously, the method is somewhat less specific than displaying actual problems to the participants (possibly depending on the wording of the items). Bong (2002) provided comparisons between measures at different levels of specificity, including subject, task and problem-specific self-efficacy in mathematics in the same studies. The results clearly demonstrated that self-efficacy at the subject, task and problem- level formed different factors, and also provided some support for the superiority of more specific self-efficacy measures (task- and problem-level) in predicting performance outcomes.

Zimmerman and Kitsansas (2005) included test taking as one facet of their Self-Efficacy for Learning Form, designed to assess students' confidence to cope with a range of academic learning or performance challenges. The authors included perceived skills in self-regulation to measure self-efficacy for taking tests (example item: *When you feel anxious during an exam and have trouble controlling information, can you relax and concentrate well enough to remember it?*).

Consistent with the literature I propose an instrument that focuses on a prospective test-taking situation. I regard problem-solving (facets one and two in my instrument: complete a certain number of problems; solve tasks of a certain challenge) as well as self-regulation (facets three and four in my instrument: concentrate; not give up) as relevant facets of the test-taking situation. By including facets related to the test-taking situation, I take into account also empirical research which has demonstrated that the relationship between self-beliefs and outcome measures is generally stronger with increasing specificity and correspondence between the measures (Valentine et al., 2004). I use items related to students' perceived capability of performing on the test, which differ from the outcomes they expect from their performance. Self-efficacy for each facet includes the "strength" of self-efficacy for each item related to the specific facet. I elaborate on the notion of strength next.

2.2.2 Strength

Strength of self-efficacy is related to the amount of individuals' confidence in their own capabilities. Bandura (1997) contends that strong self-efficacy beliefs are more resilient to change than are weakly held beliefs. Accordingly, individuals with strong positive self-efficacy are unlikely to lower their beliefs on the basis of a single adverse experience (or due to negative feedback), and individuals with strong negative self-efficacy are unlikely to

increase their expectations based on one positive performance experience alone. A majority of investigations into the relationship between self-efficacy and performance outcomes include strength of self-efficacy as a predictor of the dependent variable. Strength of self-efficacy generally predicts performance, where individuals reporting stronger self-efficacy (more confident) tend to perform better on subsequent tests in related activities. Multon et al. (1991) found an average effect size of .38 across 36 studies, while Pajares (1996b) reported that direct effects ranged from .35 to .55 in path analytic studies.

Empirical investigations indicate there are age differences in students' strength of self-efficacy, as well as in the relationship between the strength of their self-efficacy and their performance. Some researchers have found the strength of students' self-efficacy declines progressively over time (Caprara et al., 2008), and some argue a decline in self-efficacy starts during middle school (Schunk & Pajares, 2010). At the same time, the correlation between self-efficacy strength and performance seems to increase with age (Bong, 2001; Multon et al., 1991). Schunk and Pajares (2010) described how younger students are often overconfident in regard to difficult tasks, and might not fully understand task requirements. As students become older, their accuracy at self-evaluation improves, they become better at identifying relevant task features related to difficulty, and the strength of their self-efficacy, in general, declines (Schunk & Pajares, 2010). It is possible that the generally broader experience base of older students is related to this difference. Younger students might be "forced" to attend to more *domain*-specific self-efficacy when faced with unfamiliar tasks, while older students can attend to the more predictive *task*-specific self-efficacy, as they are facing more familiar tasks they are better able to assess the requirements of. As such the stronger relationship between the strength of self-efficacy and performance outcomes of older students might be related to task familiarity as well as older students' superior accuracy at self-assessment.

Gender differences have also been found in regard to strength of self-efficacy. Matsui and colleagues (1990) found a significant difference between the mathematics self-efficacy of males and females, where men reported stronger self-efficacy than women. This tendency has been reported in a number of studies on mathematics self-efficacy (Ayotola & Adedeji, 2009; Bussey & Bandura, 1999; Lloyd, Walsh, & Yailagh, 2005). For instance, Pajares and Miller (1994) found that women reported lower levels of self-efficacy in mathematics than did men, and that self-efficacy mediated gender differences found in problem-solving performance and self-concept. More recently, Schulz (2005) investigated gender differences in the PISA 2003 data. While Norway was one of the countries with the smallest absolute gender differences in mathematics literacy (effect size = -.07), effects of gender on self-efficacy (-.37), self-concept (-.43) and anxiety (.36) in Norway were all significant, with effect sizes either equal to or larger than the OECD median (-.37, -.35 and .31, respectively). Gender differences in self-efficacy strength were also found in a study by Sørlie and Söderlund (2015), where regional differences in self-efficacy were carried by differences in self-efficacy strength between the girls in the regions. It is interesting to note the presence of gender effects in self-efficacy strength in recent years, given the reportedly high levels of gender equality in Norwegian society (Ellingsæter & Leira, 2006; Ray, Gornick, & Schmitt, 2009).

Different explanations for gender effects on self-efficacy strength have been sought. Ganley and Vasilyeva (2014) argued that females become anxious about their ability in mathematics through exposure to gender stereotyping. On a challenging test, the authors found that females achieved lower test-scores and demonstrated higher levels of anxiety than males. The relationship between gender and mathematics performance was mediated by anxiety and working memory (Ganley & Vasilyeva, 2014). According to Wigfield and Tonks (2002)

gender differences in beliefs about competence have often been reported in domains related to stereotypical gender-roles, and in relation to novel tasks. Marsh, Byrne and Shavelson (1988) found that girls had higher verbal self-concept and boys had higher mathematics self-concept, after controlling for achievement, in accordance with their "differential socialization" explanation for gender differences. Pajares and Valiante (2001) investigated whether gender differences in strength of writing self-efficacy were related to gender-stereotypes. They reported that gender differences disappeared as they controlled for feminine orientation, suggesting the differences were indeed a function of gender-stereotypical beliefs rather than gender itself (Pajares & Valiante, 2001). This finding aligns with arguments that gender is a partially constructed concept, which might differ as a function of factors such as age and environmental influences, job market and cultural status (Gibbs, 2010; Penner, 2008). According to these findings, if males report stronger mathematics self-efficacy than females, the reason might be partly due to gender-stereotypical messages interpreted by female students.

Practices vary in regard to the scales applied when measuring self-efficacy strength. Some researchers advocate 0-100 (generally with 10 scale steps) (e.g. Bandura & Schunk, 1981; Hackett & Betz, 1989), some have used as few as 4 scale steps (e.g. Rimm-Kaufman et al., 2015), while others have used 5 (e.g. Midgley et al., 2000; Pajares & Miller, 1995) or 7 (e.g. Martin, 2003; Pintrich & de Groot, 1990). The anchor points used also differ between studies. Many studies use anchor points relating to certainty or confidence, such as *no confidence* to *complete confidence* (e.g. Bandura & Schunk, 1981; Hackett & Betz, 1989), or agreement, such as *strongly agree* to *strongly disagree* (e.g. Martin, 2003; Midgley et al., 2000; Pintrich & de Groot, 1990). Rimm-Kaufmann et al. (2015) adapted items from PALS to a mathematics classroom setting (example item: *I'm certain I can master the skills taught*

in math this year) and changed the anchor points from between 1 (Not at all true) and 5 (Very true) in the original measure, to between 1 (Almost never) and 4 (All the time). Scores from the 5 items were averaged for a composite score of mathematics efficacy. The change of anchor points may have changed the scale from strength-related to level-related (as "All the time" is more difficult than "Almost never"). The adaptation also makes comparisons with findings from other studies that use PALS problematic. This highlights one challenge with adapting items from a general scale.

According to Lee and Bobko (1994), strength of self-efficacy has commonly been estimated by combining the confidence ratings from all the items included in the measure. In their study they compared different ways of measuring self-efficacy, including a pure strength and a pure level measure. They asked participants to indicate whether they could perform a task at a certain level (yes/no), and to indicate the strength of their confidence on a scale from 0-10. The authors found that combining the strength scores from items that were answered "yes" led to higher convergent and predictive validities than pure strength or pure level measures.

Stajkovic and Sommer (2000) conceptualised strength of self-efficacy in relation to level (or 'magnitude'): "*Strength* of self-efficacy indicates the certainty of the subject's judgment about the magnitude" (p. 716). This is in line with the conceptualisation by Bandura (1997), that people may hold different strength self-efficacy in relation to different levels of task demands, where the strength of self-efficacy will influence behaviour and perseverance.

In this chapter I seek to disentangle the effects of facet-specificity and levels of difficulty. Like Lee and Bobko (1994), I propose a composite measure. I go beyond their study, in that I specify structural equation models where strength information is included in relation to latent

level constructs and latent facet constructs. Strength of self-efficacy is related to each individual item in the measure, which are also related to a specific facet and level of difficulty. Strength of self-efficacy is thus implicitly included in self-efficacy for each of the four facets I identified and each of the levels of difficulty (easy, medium, or hard). Furthermore, overall strength of self-efficacy can be estimated through including the confidence ratings for all items.

2.2.3 Level

Bandura (1997) emphasised that self-efficacy is not the same as people's beliefs about the skills they possess, but a measure of what they believe they can do with those skills. Accordingly, situational conditions are important, as they represent different levels of demand or impediments to success in the particular domain or task. According to Bandura "challenges may be graded in terms of level of ingenuity, exertion, accuracy, productivity, threat, or self-regulation required, just to mention a few dimensions of performance demands" (1997, p.43). Stajkovic and Sommer included levels of difficulty in their measurement of self-efficacy for naming tasks, applying the following definition: "*Magnitude* of self-efficacy measures the level of task difficulty that a person believes he or she is capable of executing." (2000, p. 716). Stajkovic and Sommer (2000) did not, however, investigate levels of self-efficacy in relation to either of the other variables in their study (causal attributions and performance at naming common uses for objects).

Bandura (1997) contended that instruments referring to a single level of task demand might fail to distinguish between individuals who differ in their self-efficacy for tasks of different levels of difficulty, thus asserting the relevance of considering different levels of task demand in measurement approaches. While many studies have investigated the aspect of generality versus specificity of self-efficacy, there seems little attention has been afforded Bandura's

proposed different perceived levels of challenge in relation to self-efficacy. This was illustrated by Multon, Brown and Lent (1991), who in their meta-analysis of 38 self-efficacy studies suggested researchers should pay more attention to relative task difficulty when measuring self-efficacy. I found only a few studies that included self-efficacy level in their measure, and similarly few studies that focused on level of self-efficacy in their analyses. Furthermore, practices for analyses of level of self-efficacy are not consistent.

One example is the Memory Self-Efficacy Questionnaire, a multidimensional self-efficacy scale devised by Berry, West, and Dennehey (1989). According to Bandura, this measure responds well to the theoretical conceptions and guidelines from self-efficacy theory and methodology. It includes items related to 10 different types of memory task, and five different levels of demand for each task. An example item is: *If I heard it twice, I could remember 12 items from a friend's grocery list of 12 items, without taking any list with me to the store.* The different levels of demand for this task include progressively easier gradations of difficulty, in regard to how many grocery items the participant can remember. For each of the different types of memory task, participants are asked to circle either "No" or "Yes". If they respond "Yes" they then rate their confidence on a scale from 10% to 100%, in 10-unit increments. For each of the 10 memory tasks the authors analysed participants' sum of "Yes" responses as self-efficacy *level,* and their average confidence across five levels as self-efficacy *strength*. The authors thus analysed self-efficacy for each memory task separately, and used intercorrelation estimates to assess task specificity post hoc.

The above measure shares similarities with the one proposed in this chapter in that it includes several types of tasks for each facet and of different levels of difficulty. At the same time, the two studies differ in conceptualisation and analyses. My chapter expands on the Berry et al.

(1989) study in that it includes both facet-specific self-efficacy (confidence across levels of difficulty; similar to the measure of self-efficacy strength in the study by Berry et al.) and self-efficacy level (confidence across four facets). In contrast, Berry et al. combined "Yes" responses across levels of difficulty to form self-efficacy level for each task.

Locke et al. (1984) included *magnitude* (meaning level) and strength of self-efficacy as two dimensions in their instrument to measure individuals' expectations about their ability to name uses for common objects. Much like the instrument developed by Berry et al. (1989), tasks at progressively more difficult levels were specified. Unlike Berry et al., Locke et al. (1984) investigated certainty ratings for varying levels of task difficulty, using strength of self-efficacy for items related to similar levels (across tasks). The authors demonstrated that their instrument could differentiate between different levels of task difficulty, and found that self-efficacy for moderate to very difficult goal levels best predicted performance. They further argued that some previous goal-setting studies had been unsuccessful in detecting a positive relationship between expectancy beliefs and performance, due to a failure to account for different levels of performance. Reanalysing the previous studies, Locke et al. (1984) found positive and significant correlations between expectancy beliefs and performance outcomes for students in the medium and difficult goal-level groups, consistent with the results from their 1984 study.

Another example is Zimmerman and Martinez-Pons (1990), who presented students with math problems of increasing difficulty (from simple arithmetic to statistics) and asked them to rate their confidence (from 0 to 100%) to solve each problem. The measures used by Locke et al. (1984) and Zimmerman and Martinez-Pons (1990) both included varying levels of difficulty. However, while Locke et al. investigated the effect of levels of task difficulty

directly, Zimmerman and Martinez-Pons (1990) combined the scores from different levels of difficulty into a composite measure, similar to pure self-efficacy strength as described by Lee and Bobko (1994) (see section 1.2.2).

I found only one study relating to mathematics self-efficacy that includes level of difficulty in the results. Chen and Zimmerman (2007) included three levels of difficulty in their analyses when investigating the relationship between levels of task difficulty and middle school students' self-efficacy in mathematics in the US and Taiwan. Students in the two countries differed in regard to their strength of self-efficacy for medium level problems, and in regard to the steepness of the decline in self-efficacy from easy to harder problems. Furthermore, students' self-efficacy for solving problems were analysed together with their subsequent performance in solving the problems. Chen and Zimmerman (2007) reported significant relationships between the difficulty level of the performance task and students' accuracy and bias in predicting own capabilities in mathematics. Bias was described as the *direction* and accuracy as the *magnitude* of students' judgment errors, when considering students' self-efficacy rating in relation to their actual performance. The authors reported that Taiwanese students' self-efficacy beliefs were more accurate and less biased than those of their US peers in relation to medium and hard level problems. However, contrary to the findings of Locke and colleagues (1984), students' self-efficacy in both countries was found to be less accurate and more biased the more difficult the math problems became (i.e. students' self-efficacy beliefs were less predictive of task performance for hard than for easy level problems). While Locke et al. (1984) specified varying levels of task demand in the self-efficacy measure a priori, the mathematics problems in Chen and Zimmerman (2007) were placed in easy, moderate, or difficult categories as determined by post-hoc assessment of students' performance. The level of difficulty was thus included in the performance measure, but not in

the self-efficacy measure. Given that self-efficacy is related to individuals' perceptions of capability, it is not certain that the easy, moderate, or difficult task levels in the study by Chen and Zimmerman (2007) are aligned with how the participants perceived the tasks before solving them. As such, the findings by Chen and Zimmerman (2007) and Locke et al. (1984) are not based on the same variables, and it is thus hard to compare and explain the contradictions that arose.

In the instrument I propose in this chapter, level of self-efficacy is related to students' perceptions of difficulty of future facets of taking a mathematics test. Levels of task difficulty are specified a priori, and included in the measurement, similar to Berry et al. (1989) and Locke et al. (1984). Level of self-efficacy includes students' certainty rating (strength of self-efficacy) for each item related to the specific level, across facets. This is similar to the conception of level of self-efficacy used by Locke et al. (1984). Levels of difficulty are also determined in relation to the national test, according to a post-hoc assessment of item difficulty, similar to Chen and Zimmerman (2007). This enables analyses into the relationship between students' self-efficacy for a certain level of difficulty problems, matched with their subsequent performance on problems of equivalent difficulty.

2.2.4 Theoretical model and research questions

Given the lack of instruments that include different levels of perceived difficulty when measuring mathematics self-efficacy, this study aims to fill the gap by proposing a multidimensional self-efficacy measure. In more detail, in this chapter I propose and investigate the structural validity of a new multidimensional instrument, which disentangles facets and levels from the strength of self-efficacy for prospective national test performance. To this end I propose a theoretical model, which is presented in Figure 2. For the 14 items measuring strength of self-efficacy I propose two factor structures to fit to the data: four test

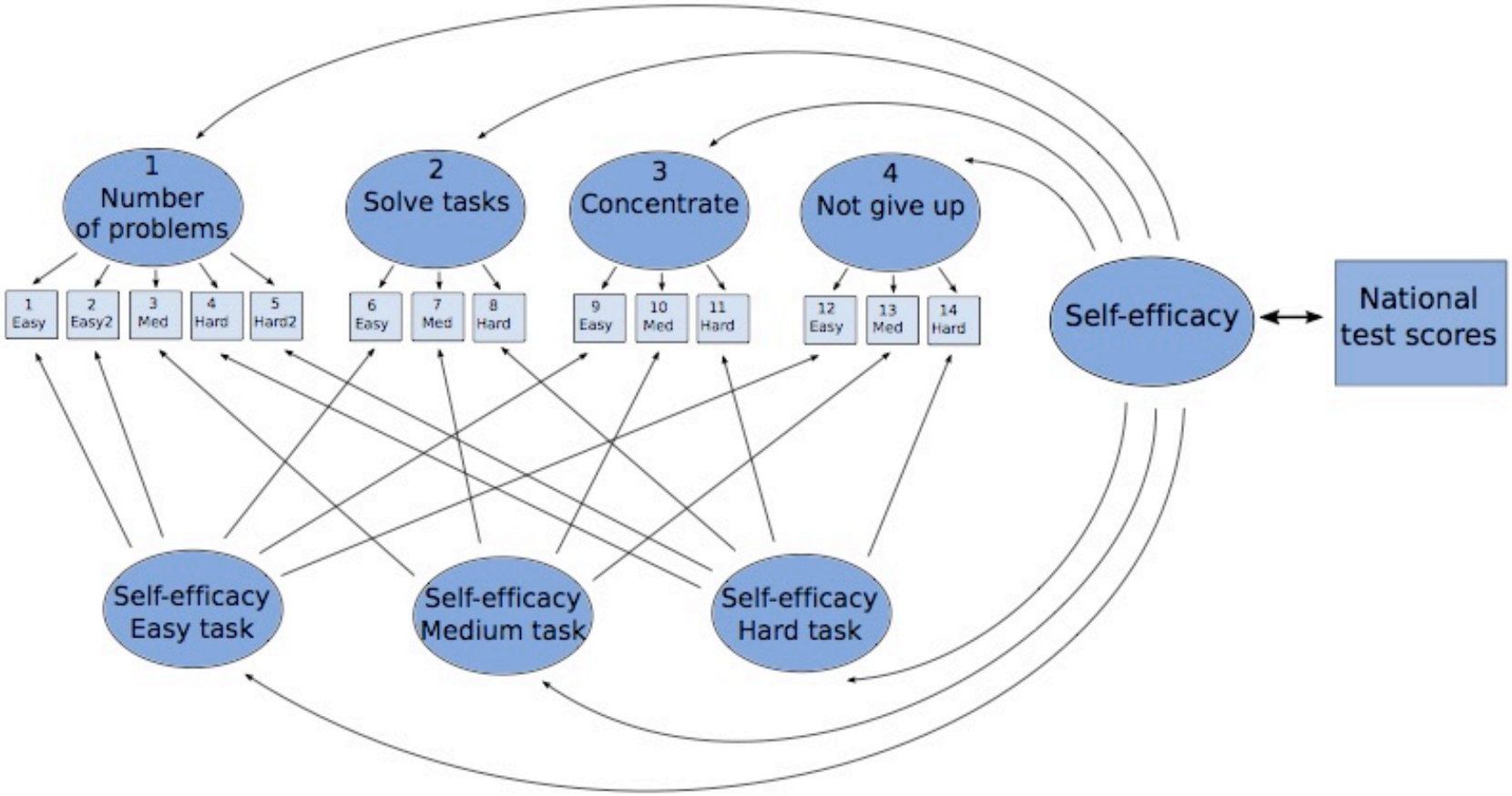
taking facets (i.e., latent constructs above indicators in Figure 2, depicting self-efficacy to (1) complete a certain number of problems, (2) solve tasks of a certain challenge, (3) concentrate, and (4) not give up for a certain amount of time (Bandura, 2006; Pajares & Graham, 1999; Zimmerman & Kitsantas, 2005)) and three perceived difficulty levels (i.e., latent constructs below indicators in Figure 2, depicting self-efficacy to solve (1) easy, (2) medium difficulty, and (3) hard tasks of the test (Berry et al., 1989; Locke et al., 1984)).

Subsequently, I hypothesise that facet-specific constructs are associated with students' national test performance, controlling for level. Likewise I hypothesise that level constructs are associated with students' national test performance, controlling for facet-specific constructs. I also hypothesize that strength could be a function of the combined scores from all 14 items, disregarding latent facets and levels, and thus an inferior estimate of self-efficacy given the multidimensional nature of the present measure. Finally, I hypothesise that students' self-efficacy for perceived easy, medium, and hard tasks are associated with their performance on easy, medium, and hard level items of the national tests. For instance, students' self-efficacy for perceived easy tasks are associated with their subsequent performance on easy test items. To conclude, the research questions I address in this chapter are:

(1) What is the structural validity of the proposed self-efficacy constructs, which include facets and levels of self-efficacy?

(2) How is facet-specific and level-specific self-efficacy associated with performance on national tests in mathematics?

(3) What is the relationship between students' self-efficacy for perceived easy, medium, and hard tasks, and the students' scores on easy, medium, and hard level items on national tests in mathematics?



Note: Items 1-14 are measured on a scale from 0 "not at all certain" to 10 "completely certain", representing strength of self-efficacy.
The four latent constructs above the indicators are test taking facets, while the three latent constructs below the indicators are levels of perceived difficulty.

Figure 2: Theoretical model for study 1

2.3 Methodology

2.3.1 Sample

Participants were 756 students (364 female) in grades 5, 8, and 9 (10, 13 and 14 years old, respectively) in Norwegian state schools (see Table 1). Students from these grades were selected given that national tests in numeracy are held for all 5th, 8th and 9th graders in Norway each year. The test is the same for students in grade 8 and grade 9. Study 1 of the thesis is linked with a larger Norwegian project (Langfeldt, 2015), and the participants were selected according to specific, joint selection criteria. Participating schools were selected from municipalities within four regions in Norway. The schools were purposively selected, according to available information regarding socio-economic factors (data for this was available at the municipal level). Taking this information into account, there was an aim to include both schools where students had performed above what might be expected and schools where students had performed below what might be expected, for the last three years. I included schools with between 150 and 400 students, which is a medium size school in the Norwegian context. Education advisors with the regional educational directorates were asked to verify the choices of schools, in an aim to select a range of schools in each region, and to enable contextual interpretations of the findings. The current sample is thus not randomized, but based on given selection criteria. I argue that the primary and secondary student populations within the regions included is representative to the primary and secondary student population in Norway in general (for further evidence of this, see Tables 4-7 in section 1.4.3, *Instruments*). The Norwegian school system, where more than 96% of students attend state schools (Statistics Norway, 2017) with little or no forms of tracking or ability grouping, provides an advantageous setting in this respect, in regard to the generalizability of findings from my sample to the population, in this case Norwegian school students in grades 5, 8, and 9.

The data collection was conducted in two subsequent years, 2012 and 2013. In 2012, a total number of 14 schools participated. In 2013, 10 of the schools partaking in 2012 agreed to participate again, and an additional 13 schools also accepted the invitation to partake. This resulted in a total number of 43 classes from 27 schools, over the two years. The total number of students attending these classes was 1009, where 851 (84%) consented to partake in the study (in year one 393 out of 489 students – i.e. 80% - consented; in year two 458 out of 520 students – i.e. 88% - consented). Students from five grade eight classes in 2012 (n=95 students) participated also the subsequent year. For research questions one and two, the repeat students are only included once, with their 2012 data. This results in a sample of 756 *unique* participants, from 38 classes. For research question three all participants (including repeat participants) are included in the sample. This results in a sample of 851 students, from 43 classes. Table 1 shows participants according to grade, year, and gender. The numbers of unique participants in grade 9 (i.e., participants that had not already taken part in grade 8) are presented in bold type. In the grade 5 and grade 8 samples all participants are 'unique' – that is, they had not previously taken part in the study.

Table 1: Participants study 1

Unique participants (RQ1 & RQ2), N = 756; all participants (RQ3), N = 851

Year	Grade 5 (n=363)		Grade 8 (n=228)		Grade 9 (n=260 / **165**)	
	Female	Male	Female	Male	Female	Male
2012	82	83	101	127	-	-
2013	104	94	-	-	121 **(77)**	139 **(88)**
Total	186	177	101	127	121 **(77)**	139 **(88)**

Note: For grade 9, the unique participants are presented in **bold**

For research questions one and two, the participant samples from the two years of data collection were merged, so that the grade 5 sample in 2012 was combined with the grade 5 sample in 2013, and the grade 8 sample in 2012 was combined with the grade 9 sample in 2013. The reason for merging the samples was to increase the sample size to enable statistical

analyses, and to be able to present the findings coherently. However, there can also be disadvantages associated with combining different samples. One concern is whether the samples differ in regard to their scores on the variables of interest, and whether these variations influence the research findings. Item descriptives from grades 5, 8, and 9 in year 2012 and 2013 are presented in Tables 2 and 3, in section 2.3.3. Table 2 indicates there are no systematic differences between the two grade 5 samples.

In regard to the grade 8 and grade 9 samples, there is reason to expect grade 9 students' national test scores to differ (be higher) than grade 8 students' scores, due to their extra year of school as well as maturity levels. This is confirmed in the descriptives presented in Table 3. In regard to self-efficacy, grade 9 students' scores are comparable to the scores of the grade 8 students. It is possible that the relationship between scores on the self-efficacy constructs and national tests are higher for the grade 9 than the grade 8 sample. This is a potential disadvantage to combining grades 8 and 9 in research questions one and two, as any grade level variations to the self-efficacy / national test score relationship will be "hidden" by the mean estimates.

The response rates varied greatly between the participating classes. The response rate depended on how many students consented, but was also affected by things like absenteeism on the day of the questionnaire or the day of the national test. There were some students who took part in the questionnaire, but were absent from the national tests, or vice versa. In two classes the response rate was as low as 25% and 27%, while several classes had 100% response rates. Response rates may be related to classroom or home factors, and low response rates can influence research results if the participants differ systematically from the ones who did not partake (Tabachnick & Fidell, 2013). I organized the participating classes in three

categories, according to their response rates. Category three included classes with higher than 74% response rates (26 classes, n= 575), category two included classes with between 44 - 65% response rates (9 classes, n= 162), while category one included classes with response rates below 33% (3 classes, n= 19). All the classes fitted within one of these three categories. I used these categories as filters to run the analyses with and without the low-response rate classes, to check whether this would influence the model fit indices or the parameter estimates. The results from these checks are elaborated upon in section 2.4.3. In short, while the parameter estimates are not identical when different filters for response rates are used, including the lower response classes does not influence which models are best-fitting. Subsequently I argue that I cannot justify excluding the low response classes, and I decided to include both high and low response classes in my investigations.

2.3.2 Ethical considerations

For any given study it is important to consider the potential benefit of findings versus the inconvenience to participants and potential harmful effects of the study. This discussion of ethical considerations relate to studies 1 and 2 of the National Test Phase of this thesis. This phase involves children under the age of 18, thus parental consent was sought. Participant information letters (see Appendix D) were distributed to all students and their parents via their teachers. These letters informed the participants of the name of the study ("Mestringsforventninger") as well as the broad purpose of the study (understand how to facilitate teaching in schools to improve student attainment in mathematics). Participants were informed what was involved if they chose to take part (filling out a questionnaire at school), what kinds of questions the students would be asked in the questionnaire, and that student responses would be analysed in relation to their national test results. The students were informed that the teacher would also be filling out a questionnaire, and answer questions about the students. The information letters also informed participants about the manner in

which the data would be handled, emphasised that participation was voluntary, and that the students could withdraw at any time, without stating any particular reason.

The teachers received information about the study through an information letter sent to each school (Appendix E). These letters informed the teachers of the purpose of the study, what was required of them if they chose to partake, and that participation was voluntary. In addition, the letters encouraged the teachers to contact me as principal investigator if they had any queries.

Information letters were sent out to student and teacher participants in both 2012 and 2013. The letters included similar information each year, with minor adjustments referring to the stage of the study. Furthermore, a short section was added to the teacher information letter in 2013 (the second year of data collection), where I asked each teacher to send me an email to establish direct contact. This was considered important in regard to avoiding misunderstandings and building a trusting relationship with the teachers. The decision to add this section was made partially on the basis of some misunderstandings and misconceptions discovered during the first year of data collection. For instance, in 2012 one teacher who had misunderstood my role in relation to the national tests (she did not agree with the tests), expressed surprise when probed why she had not contacted me directly about her concerns, stating *I saw that you wrote I could ask you questions* <in the information letter >*, but I did not think it really applied to* ***me***. I have no reason to suspect that the response rate or the final results were affected due to misunderstandings or misconceptions in 2012. However, the more direct communication with the teachers during the second year of data collection might have ensured teachers had a more positive experience of taking part of the study, and helped to avoid misunderstandings and complications.

The impact on each of the participants regarding their participation in the study was not very large (851 students x maximum 1 hour, 43 teachers x approximately 3 hours), and it was deemed that students' experiences during data collection were not very dissimilar to their everyday experiences in school. However, there were many participants, and the data collected was of a somewhat sensitive character (individual test results linked with questionnaire responses). Accordingly, care has been taken in the handling and storage of this data. Identifying information is kept on separate files to the test results and questionnaire responses, and all manifestations of the data (physical or electronic) are stored in password-protected files or locked cabinets in the offices of Karin Street (at the University college of Sogn & Fjordane or in Oxford) or Dr Lars Malmberg (Oxford).

With a view to ensure that the participants benefit from the study directly, all participating schools have been promised a final report about the key findings emerging from the study. I will write this report considering the mathematics teachers as audience, and provide classroom examples where relevant. In addition, together with a Norwegian colleague I have already visited nine of the schools, giving talks about self-efficacy. I consider the benefits of the study, e.g. knowledge about self-efficacy for participating schools and teachers, and knowledge regarding self-efficacy in the mathematics classroom for the research community, to outweigh any negative impact on the participants in regard to inconvenience and resources spent. Ethical guidelines are followed, and ethical approval has been sought and obtained in both Norway and Oxford (see Appendix F) in order to minimize the risk of harm to participants.

2.3.3 Measures

The data collected consisted of questionnaires measuring students' self-efficacy, and scores from national tests in mathematics.

Self-efficacy

Students' self-efficacy was assessed with the Self-Efficacy Gradations of Difficulty Questionnaire (see Appendix I) that I developed for the present study. This new instrument includes 14 items, where each item is related to one of four facets[1] concerning problem solving and self-regulation on national tests in mathematics, as well as one of three levels of difficulty within each facet. For each item participants were asked to indicate their confidence on an 11-point scale from 0 "not at all certain" to 10 "highly certain", which represents strength of self-efficacy.

Instrument development

I developed the Self-Efficacy Gradations of Difficulty Questionnaire following guidelines made available by Bandura (2006). I wished to create a measure that was brief and could meaningfully relate to taking national tests, including some representative tasks. Thus I included in the measure facets related to problem solving, which is commonly investigated in the mathematics self-efficacy literature, but also self-regulation, given the theoretical link between self-regulation and self-efficacy (Zimmerman, 2000), and the importance of self-regulation to performance (Throndsen, 2011). While the included facets were considered relevant and representative, the proposed measure is not exhaustive, as I also had to consider the length of the questionnaire in regard to the age of the respondents. Additional or

[1] A 5th facet, "Grade", was originally tested, but not deemed adequate. Examination of individual student responses indicated that the items were confusing.

alternative facets could have been included in the questionnaire, and it may be meaningful to expand the measure to include other facets in future studies, such as for instance "to apply relevant strategies during problem solving" and "to think positively".

To assess face-validity, the instrument was piloted with six students in 4th grade of varying mathematics proficiency. First, I explained the questionnaire, and completed one example item together with the students (self-efficacy for lifting an object of increasing weight), using the classroom blackboard. The 4th grade students were then told to take their time on the questionnaire, to answer truthfully, and that their answers would not be shown to anyone, such as their teachers. I was present while the students worked on the questionnaires, and encouraged them to ask questions. When they finished, I spoke with each student individually, asking them whether they had found any of the items difficult to understand, and to "talk through" a couple of items on the questionnaire. The students were asked to explain what the item meant in their own words, and explain why they answered as they did. These interviews indicated the students understood the items on the self-efficacy questionnaire, and that they understood how to indicate the strength of their self-efficacy for each item. One of the students took a long time (longer than 30 minutes) on the questionnaire, which indicated to me that some students might require assistance when filling out the questionnaire, with respect to their reading skills. The questionnaires included other measures not used in the present study, for instance the measure of ability conceptions developed by Dweck (2008). As a result of the long completion time during the pilot, items from some of the other measures were deleted, in order to shorten the length of the questionnaire.

After the first round of data collection (in 2012), I conducted interviews with eight student participants, four in grade 8 and four in grade 5, in order to check whether the students had

interpreted the items in the intended way. The teachers assisted me in finding students who were happy to chat with me about the questionnaires. The outcomes of these interviews indicated the students' interpretations were generally aligned with the intended meaning of the items. One interesting finding from an interview with a male student in grade 8 was that the student reported feeling more confident he could complete hard tasks, than easy tasks. The student described himself as high achieving, and explained he might not concentrate enough if he perceived of the task as "too easy". I did not find this unexpected response pattern to be reason to change the questionnaire, however I took a note of his explanation in regard to data analyses.

Although the questionnaire was piloted and I also conducted student interviews to assess their interpretations of the items, I later discovered that a number of students had misinterpreted the meaning of facet (5) to *achieve a certain grade*. Initially, five test-taking facets were included. These were to (1) complete a certain number of problems, (2) solve tasks of a certain challenge, (3) concentrate, (4) not give up for a certain amount of time, and (5) achieve a certain grade. The facets were included based on previous self-efficacy instruments developed by Bandura (2006), Zimmerman and Kitsansas (2006), and Betz and Hackett (1983). When I considered the item descriptives (including items related to the fifth facet) these indicated a different tendency of responses to item 15 *During the national tests this year I can achieve at least grade 1* (all samples) and item 16 *During the national tests this year I can achieve at least grade 2* (grade 8/9 sample only), compared to other items (see Appendix G). These two items, both associated with the "achieve a certain grade" facet and the easy level of difficulty, had lower mean estimates and higher standard deviations than other items associated with the easy construct. This is contrary to what one might expect, given that all students will definitely achieve the minimum grade (the minimum grade is assigned if

students get from 0-12 points in grade 5, and from 0-18 points in grade 8/9). Accordingly, it would be reasonable to expect high means and little variance (and thus standard deviations) in relation to items 15 and 16.

The grade 8/9 correlation matrix (see Appendix H) also indicated issues with items 15 and 16. Item 15 was correlated with the least other items in the measure (10 out of 18 correlations non-significant), followed by item 16 (7 out of 18 correlations non-significant). At the same time the two items were highly intercorrelated (r=.959).

I went on to inspect the individual response patterns of grade 8/9 students in the raw data. This pointed to a problem with the fifth facet, where some student responses to this facet were inconsistent with the intended interpretation of the items. For instance, students who indicated they were confident they would achieve a high grade (high scores on items 18 *During the national tests this year I can achieve at least grade 4* or 19 *During the national tests this year I can achieve at least grade 5*) indicated they were not confident they would achieve at least the minimum grade (low scores on items 15 and 16). These trends indicated the items were ambiguously worded, or that some students failed to take notice of the wording "at least". It seems likely some students responded to the *probability* they would get a certain grade rather than their confidence in achieving *at least* a certain grade (the intended meaning of the items). This seeming inconsistency was observed for as many as 59 out of 373 grade 8/9 students (15.8%). Given these findings, grade 5 students' raw data were also inspected. Similar tendencies were found: scores from 39 of 337 students (11.6%) were inconsistent with the intended meaning of the items associated with the facet "achieve a certain grade".

While the descriptive data and the grade 8/9 correlation matrix indicated problems with items 15 and 16 in particular, interpretation of responses to all items associated with this facet is problematic if a high number of students misinterpreted the wording of the items. It can also not be said for sure which students interpreted the items differently, thus deleting individual student responses seems unreasonable. For this reason all items associated with the fifth facet, "achieve a certain grade" were excluded from further analyses in the case of both grade 5 and grade 8/9.

Descriptives for the remaining 14 items are presented in Tables 2 and 3. When I inspected these descriptives items 1, 6 and 9 were severely skewed and peaked (Kline, 2011, p. 63). As most students are likely to be able to complete the easy tasks, it is reasonable that items associated with this level are non-normally distributed. In order to deal with this data, I used the robust maximum likelihood estimator in Mplus for analyses (see section 2.3.5).

Table 2: Item descriptives, grade 5

	Item	Grade 5, 2012						Grade 5, 2013					
		n	M	SD	Range	Skew	Kurt	n	M	SD	Range	Skew	Kurt
1	No. of problems_easy	151	10,75	1,19	10	-5,84*	-5,84	186	10,61	1,45	10	-4,59*	23,22*
2	No. of problems _easy2	151	9,95	2,01	10	-1,99	-1,99	186	9,98	2,03	10	-2,24	4,77
3	No. of problems _med	151	8,50	2,77	10	-0,77	**-0,77**	186	8,26	2,71	10	-0,70	**-0,56**
4	No. of problems _hard	151	6,73	3,09	10	**-0,14**	-0,14	186	5,65	2,98	10	**0,15**	-0,90
5	No. of problems _hard2	150	4,05	3,15	10	0,65	0,65	186	3,92	2,79	10	0,67	**-0,52**
6	Solve tasks_easy	151	10,40	1,36	10	-3,82*	-3,82	186	9,92	1,86	10	-2,40	6,49
7	Solve tasks_med	151	7,81	2,07	8	**0,09**	0,09	186	7,77	2,14	10	-0,36	**-0,20**
8	Solve tasks_hard	151	4,81	2,70	10	**0,27**	0,27	186	5,31	2,75	10	**0,09**	-0,84
9	Concentrate _easy	151	10,38	1,93	10	-3,69*	-3,69	186	10,23	2,20	10	-3,14*	9,17
10	Concentrate _med	151	9,48	2,23	10	-1,34	-1,34	186	9,31	2,53	10	-1,69	2,27
11	Concentrate _hard	151	7,46	3,22	10	-0,46	-0,46	186	7,8	2,97	10	-0,78	**-0,38**
12	Not give up _easy	151	9,42	2,47	10	-1,59	-1,59	186	9,09	2,70	10	-1,38	1,15
13	Not give up _med	151	8,01	2,69	10	-0,46	-0,46	186	8,28	2,82	10	-0,76	**-0,51**
14	Not give up _hard	151	6,74	3,34	10	**-0,22**	-0,22	186	7,22	3,42	10	-0,46	-1,17
15	National test score	163	27,58	9,15	39	**-0,21**	-0,21	188	27,65	8,17	40	-0,37	**-0,10**

Note: z-score is larger than 1.65 for all items, *except* those in **bold**

*Absolute value higher than recommended cut-off, e.g. Skew Index >3, Kurtosis Index >10 (Kline, 2011, p.63)

Table 3: Item descriptives, grades 8 and 9

	Item	Grade 8, 2012						Grade 9, 2013					
		n	M	SD	Range	Skew	Kurt	n	M	SD	Range	Skew	Kurt
1	No. of problems_easy	221	10,77	0,99	9	-5,70*	37,11*	239	10,73	1,21	10	-5,51*	33,35*
2	No. of problems _easy2	221	10,35	1,40	9	-2,65	7,75	239	10,22	1,77	10	-2,74	7,80
3	No. of problems _med	222	8,90	2,25	9	-0,78	**-0,39**	239	8,92	2,52	10	-1,13	**0,51**
4	No. of problems _hard	222	6,50	2,64	10	**-0,01**	-0,78	239	6,29	2,78	10	**-0,02**	-0,87
5	No. of problems _hard2	220	3,76	2,85	10	0,87	**-0,29**	239	3,55	2,94	10	0,98	**-0,17**
6	Solve tasks_easy	221	10,11	1,44	9	-2,38	7,04	239	10,14	1,55	8	-2,38	5,69
7	Solve tasks_med	222	7,88	2,04	10	-0,72	0,69	239	8,14	2,14	10	-0,66	**0,52**
8	Solve tasks_hard	221	5,34	2,46	10	**0,03**	-0,78	239	5,13	2,73	10	**0,18**	-0,91
9	Concentrate _easy	221	10,29	2,04	10	-3,17*	9,69	239	10,32	1,84	10	-3,25*	11,15*
10	Concentrate _med	221	9,45	2,27	10	-1,63	2,09	239	9,45	2,23	10	-1,43	1,39
11	Concentrate _hard	222	7,47	3,01	10	-0,62	-0,61	239	7,31	3,17	10	-0,53	-0,90
12	Not give up _easy	221	9,23	2,57	10	-1,54	1,63	239	9,48	2,34	10	-1,76	2,71
13	Not give up _med	222	8,17	2,57	10	-0,83	**0,13**	239	8,43	2,38	10	-0,81	**0,29**
14	Not give up _hard	221	6,87	2,84	10	**-0,23**	-0,77	239	6,66	3,16	10	**-0,20**	-1,12
15	National test score	213	28,56	11,44	48	**0,10**	-0,85	252	34,54	11,18	53	-0,37	**-0,46**

Note: z-score is larger than 1.65 for all items, *except* those in **bold**

*Absolute value higher than recommended cut-off, e.g. Skew Index >3, Kurtosis Index >10 (Kline, 2011, p.63)

National test scores

National test scores in numeracy (Norwegian Directorate for Education, 2016) were used as the performance measure and dependent variable. The national tests were developed for the Norwegian Directorate for Education to assess to what degree Norwegian schools are successful in developing students' competencies in Norwegian, English and numeracy. What is referred to as "numeracy" is essentially the same as what researchers frequently refer to as mathematics. Students in 5th, 8th, and 9th grade sit computer-based national tests in mathematics each autumn (the beginning of the school year). The tests for 5th grade are based on the national goals for 4th grade, while students sit the same test in 8th and 9th grade, based on the national goals for 7th grade. Students in 5th grade are presented with 45 problems while students in 8th and 9th grades have 58 problems to solve. Each problem is scored as either 1 (correct) or 0 (incorrect). To enable meaningful analyses, I used students' absolute scores on the test (score out of 45 or 58 possible) rather than their resulting grade, which is dependent upon the national average each year. Students' national test scores are presented in Tables 2 and 3. National test raw scores were linked to the questionnaire data by unique identifiers.

In order to check whether my sample is representative of the population, I used what is termed the "p-value" for each test item, obtained from the national directorate for education. The p-value represents item difficulty, and is the proportion of students in the national sample that solved the task correctly (i.e., got the score "1" rather than "0"). I calculated the corresponding item difficulty values in my sample, which enabled me to compare my sample with the national sample, in regard to students' performances on the national tests. Comparisons between my sample and the national sample are displayed in Tables 4 to 7. In these tables the test items are ranked according to item difficulty in my sample, e.g. the

percentage of students in my study that got each item correct. Item difficulty is indicated by colour (easy, medium, and hard = green, orange, and red), and lines in the tables indicate cut-offs between the groups. Descriptives for students' scores on easy, medium, and hard items are presented in section 2.4.4. As can be seen in Tables 4 to 7, for most of the items the difficulty values are comparable. This is demonstrated in the mean difficulty values, which differ by a maximum of 5 percentage points.

Table 4: Comparison of national test performance (grade 5, 2012)

Rank-order: Item difficulty (easy-hard)	Item number	Difficulty my sample (n=165)	Difficulty national sample (n=18.412)
1	item 4	94 %	91 %
2	item 12	90 %	91 %
3	item 3	90 %	82 %
4	item 10	83 %	84 %
5	item 8	83 %	68 %
6	item 16	80 %	77 %
7	item 42	79 %	81 %
8	item 27	79 %	74 %
9	item 18	79 %	78 %
10	item 5	78 %	78 %
11	item 9	77 %	72 %
12	item 2	77 %	93 %
13	item 14	75 %	72 %
14	item 33	74 %	71 %
15	item 1	74 %	89 %
16	item 29	72 %	63 %
17	item 22	71 %	67 %
18	item 40	69 %	71 %
19	item 20	69 %	72 %
20	item 6	67 %	61 %
21	item 31	66 %	56 %
22	item 32	63 %	62 %
23	item 26	63 %	60 %
24	item 21	63 %	60 %
25	item 25	60 %	48 %
26	item 24	56 %	49 %
27	item 35	55 %	53 %
28	item 23	55 %	57 %
29	item 13	53 %	48 %
30	item 30	52 %	56 %
31	item 41	51 %	46 %
32	item 43	50 %	56 %
33	item 36	48 %	43 %
34	item 19	45 %	47 %
35	item 15	45 %	45 %
36	item 45	44 %	30 %
37	item 38	44 %	52 %
38	item 7	42 %	42 %
39	item 39	40 %	48 %
40	item 37	39 %	35 %
41	item 34	38 %	40 %
42	item 44	36 %	20 %
43	item 17	36 %	39 %
44	item 11	36 %	32 %
45	item 28	18 %	14 %
Mean difficulty value:		61 %	59 %

Table 5: Comparison of national test performance (grade 5, 2013)

Rank-order: Item difficulty (easy-hard)	Item number	Difficulty my sample (n=198)	Difficulty national sample (n=17.926)
1	item 10	91 %	84 %
2	item 1	88 %	80 %
3	item 14	86 %	79 %
4	item 33	83 %	78 %
5	item 30	80 %	74 %
6	item 20	80 %	78 %
7	item 4	80 %	69 %
8	item 22	78 %	70 %
9	item 6	78 %	63 %
10	item 31	77 %	75 %
11	item 29	76 %	75 %
12	item 3	76 %	88 %
13	item 16	74 %	62 %
14	item 8	72 %	35 %
15	item 21	69 %	67 %
16	item 43	68 %	47 %
17	item 18	68 %	62 %
18	item 5	68 %	80 %
19	item 12	66 %	62 %
20	item 9	66 %	44 %
21	item 24	65 %	67 %
22	item 2	64 %	80 %
23	item 32	64 %	57 %
24	item 27	64 %	56 %
25	item 40	62 %	80 %
26	item 35	61 %	59 %
27	item 41	57 %	57 %
28	item 38	57 %	55 %
29	item 7	57 %	56 %
30	item 44	54 %	27 %
31	item 26	52 %	49 %
32	item 11	52 %	42 %
33	item 45	51 %	46 %
34	item 34	49 %	44 %
35	item 15	48 %	44 %
36	item 17	47 %	39 %
37	item 42	46 %	41 %
38	item 23	46 %	43 %
39	item 19	41 %	40 %
40	item 37	39 %	35 %
41	item 13	39 %	36 %
42	item 39	37 %	12 %
43	item 25	36 %	37 %
44	item 36	28 %	45 %
45	item 28	18 %	22 %
Mean difficulty value:		61 %	56 %

Table 6: Comparison of national test performance (grade 8, 2012)

Rank-order: Item difficulty (easy-hard)	Item number	Difficulty my sample (n=232)	Difficulty national sample (n=19.609)
1	item 20	91 %	91 %
2	item 16	82 %	75 %
3	item 2	77 %	92 %
4	item 25	76 %	76 %
5	item 23	76 %	76 %
6	item 24	75 %	76 %
7	item 8	75 %	58 %
8	item 5	75 %	71 %
9	item 39	74 %	70 %
10	item 21	74 %	77 %
11	item 12	70 %	72 %
12	item 33	68 %	66 %
13	item 47	67 %	65 %
14	item 6	67 %	86 %
15	item 1	67 %	67 %
16	item 46	66 %	71 %
17	item 45	66 %	61 %
18	item 35	66 %	73 %
19	item 17	65 %	63 %
20	item 11	63 %	57 %
21	item 7	62 %	55 %
22	item 18	61 %	58 %
23	item 9	59 %	57 %
24	item 19	57 %	64 %
25	item 52	56 %	54 %
26	item 41	55 %	57 %
27	item 50	53 %	36 %
28	item 49	53 %	41 %
29	item 15	52 %	51 %
30	item 54	50 %	37 %
31	item 42	49 %	50 %
32	item 3	49 %	43 %
33	item 30	48 %	47 %
34	item 27	47 %	43 %
35	item 31	43 %	49 %
36	item 51	42 %	52 %
37	item 44	41 %	39 %
38	item 4	40 %	49 %
39	item 53	39 %	27 %
40	item 14	39 %	40 %
41	item 56	38 %	62 %
42	item 10	34 %	20 %
43	item 55	33 %	63 %
44	item 37	33 %	39 %
45	item 48	29 %	36 %
46	item 32	29 %	28 %

47	item 26	27 %	24 %
48	item 40	25 %	28 %
49	item 22	24 %	23 %
50	item 38	22 %	24 %
51	item 29	22 %	26 %
52	item 13	22 %	24 %
53	item 58	17 %	2 %
54	item 57	17 %	13 %
55	item 34	13 %	18 %
56	item 36	10 %	8 %
57	item 28	10 %	11 %
58	item 43	8 %	11 %
Mean difficulty value:		49 %	49 %

Table 7: Comparison of national test performance (grade 9, 2013)

Rank-order: Item difficulty (easy-hard)	Item number	Difficulty my sample (n=260)	Difficulty national sample (n=19.805)
1	item 24	90 %	88 %
2	item 20	87 %	83 %
3	item 6	86 %	78 %
4	item 1	86 %	86 %
5	item 41	85 %	79 %
6	item 40	81 %	79 %
7	item 39	81 %	77 %
8	item 2	81 %	92 %
9	item 45	80 %	76 %
10	item 25	80 %	75 %
11	item 5	78 %	60 %
12	item 21	75 %	69 %
13	item 15	75 %	75 %
14	item 31	74 %	73 %
15	item 47	73 %	70 %
16	item 22	73 %	69 %
17	item 17	73 %	72 %
18	item 16	73 %	70 %
19	item 11	73 %	75 %
20	item 7	72 %	62 %
21	item 18	71 %	71 %
22	item 12	71 %	73 %
23	item 14	70 %	63 %
24	item 9	69 %	73 %
25	item 32	68 %	60 %
26	item 4	68 %	67 %
27	item 46	67 %	69 %
28	item 8	65 %	75 %
29	item 33	64 %	60 %
30	item 37	63 %	58 %
31	item 19	59 %	57 %
32	item 23	58 %	55 %
33	item 52	56 %	55 %
34	item 50	56 %	45 %
35	item 3	56 %	74 %
36	item 30	54 %	50 %
37	item 27	54 %	58 %
38	item 13	54 %	46 %
39	item 56	52 %	69 %
40	item 53	51 %	35 %
41	item 10	51 %	23 %
42	item 49	48 %	44 %
43	item 55	46 %	58 %
44	item 44	44 %	39 %
45	item 51	40 %	32 %
46	item 48	40 %	38 %

47	item 34	40 %	42 %
48	item 29	40 %	35 %
49	item 54	39 %	37 %
50	item 35	37 %	35 %
51	item 57	36 %	21 %
52	item 38	35 %	31 %
53	item 26	32 %	37 %
54	item 58	27 %	12 %
55	item 42	27 %	25 %
56	item 28	26 %	28 %
57	item 43	24 %	21 %
58	item 36	18 %	16 %
Mean difficulty value:		60 %	57 %

2.3.4 Empirical models

In order to test multidimensionality of my measure, I specified a set of logical models, ranging from simple (e.g., correlated latent constructs) to complex (e.g., correlated uniquenesses, higher-order constructs) (see Figures 3 to 13). The specifications were based on the current conceptualisation of self-efficacy as well as existing approaches to their measurement, as reviewed in section 2.2.

In my theoretical model (Model 7) latent constructs (commonly referred to as Latent Method Factors, LMF) represent both levels and facets, based on the logic of the multi-trait multi-method (MTMM) model (Campbell & Fiske, 1959). An alternative to including latent constructs is to include correlated uniquenesses (CUs), letting dimensions or levels (respectively) represent the method effect these strategies are commonly used to address. According to Marsh and colleagues (Marsh, Scalas, & Nagengast, 2010), one advantage of specifying CUs is the models tend to converge, while an advantage of specifying LMFs is the researcher can summarize the effects of the latent factors and relate them to other factors in the model. It is debated whether one approach is preferable to the other (Chiorri, 2014). Given LMF types of models are especially prone to non-convergence or improper solutions (Marsh, Scalas, & Nagengast, 2010), in addition to my theoretical model I also specified models with correlated uniquenesses (CUs) representing levels or facets. Both strategies enable me to consider both facets and levels of mathematics self-efficacy. Below I describe each of my empirical models.

Baseline model (Model 1; Figure 3)

The most parsimonious, baseline model, is the *unidimensional* one, where all 14 items were specified to load on a single self-efficacy construct, representing overall strength of self-

efficacy. While model 1 is useful for comparison purposes, it is unlikely to fit the data well as it does not take into account the facets or the levels hypothesized.

Models with correlated facets or levels (Models 2 and 3; Figures 4 and 6)

Following on from the unidimensional model, competing hypotheses can be formulated according to whether facets or associated levels of difficulty are thought to influence students' self-efficacy more. These can be specified in *correlated facets* models and *correlated levels* models, where item scores are predicted by a set of factors related either to the different test-taking aspects (facets: Model 2) or the varying degrees of perceived difficulty of the items (levels: Model 3). As discussed in the review of existing approaches to measurement of self-efficacy, previous research points to the subject-specific nature of self-efficacy. This supports a model taking into consideration the differences between the 4 facets specified (Model 2).

Correlated levels models have on the other hand seemingly not previously been tested. In the few exceptions known (see section 2.2.3), SEM has not been applied. In the study by Berry, West, and Dennehey (1989), for instance, scores for memory self-efficacy were aggregated from different levels of difficulty for each type of task. In the current context, a correlated levels model would rely on a theoretical assumption that students' self-efficacy is first and foremost related to the degree of perceived difficulty of each of the facets in question (Model 3).

Correlated facets /correlated levels models with CUs (Models 2b and 3b; Figures 5 and 7)

Neither Model 2 nor Model 3 take into account the multidimensionality of the present self-efficacy measure. According to Bandura's theory self-efficacy is related to specificity, level

and strength, and it is theoretically plausible that both facets and levels of difficulty influence students' self-efficacy.

A model with the facets as latent factors and the levels of challenge as a-priori specified error-terms (Model 2b) build on the theoretical assumptions of Model 2, where the different facets are the most salient aspect of self-efficacy, while different levels of difficulty of the items captures some common variance. Model 2b was thus specified as an alternative to model 2. Similarly, a model with the levels of difficulty as latent factors and the facets as a-priori specified error terms (Model 3b) relies on the same theoretical arguments as Model 3, in regard to the influence of degree of perceived difficulty on students' self-efficacy. Model 3b was specified as an alternative to Model 3.

Second-order models with latent facet / level factors (Models 4 and 5; Figures 8 and 10)

As alternative to the first-order models, I specified higher order models in which an underlying self-efficacy construct predicts four facets (Model 4), and three levels (Model 5). Second-order models are often proposed and supported in the self-efficacy literature, for instance if self-efficacy is seen as subject-specific (and thus differs between different subjects), but load on a single underlying self-efficacy factor (e.g. Bong 2001). In this study, such models rely on the assumption that students' self-efficacy beliefs differ substantially between different facets (Model 4) or different levels (Model 5) yet reflect a higher-order construct (self-efficacy).

Second-order models with CUs (Models 4b and 5b; Figures 9 and 11)

As in the case of Models 2 and 3, I specified multidimensional alternatives to Models 4 and 5. Models 4b and 5b include a priori determined correlated uniquenesses for levels, and facets, respectively.

Latent constructs / method factor (LMF) models (Models 6 and 7; Figures 12 and 13)

An alternative approach to include both facets and levels of difficulty are Latent Method Factor (LMF) models. In such models both facets and levels of difficulty are included as latent, correlated factors, independent of one-another. In Model 6, I specified four facets and three levels as latent, correlated factors, where facets and levels were presumed independent of one-another. In Model 7, my theoretical model, I specified self-efficacy as a higher order construct predicting four latent facet constructs and three latent level constructs.

After testing the fit of my specified models, I investigated associations between students' self-efficacy and national test scores. For this, the best-fitting of the specified models was used as the measurement part of a full structural equation model including national test scores as the correlate. For comparison purposes, the most viable alternative measurement models (from a theoretical as well as fit indices perspective) were also tested.

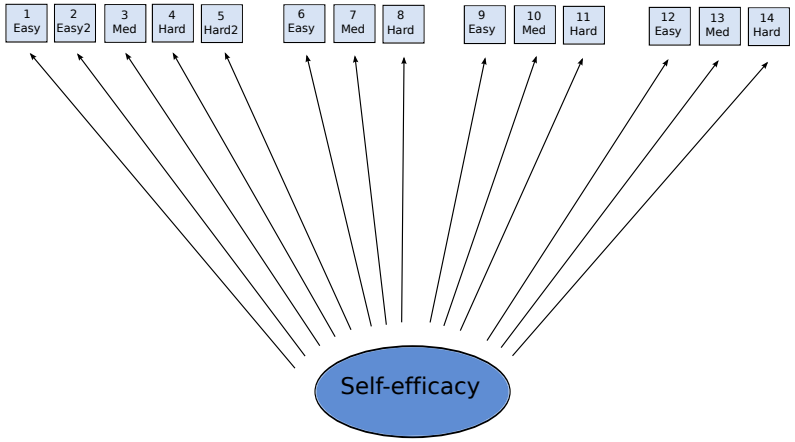


Figure 3: Unidimensional (Model 1)

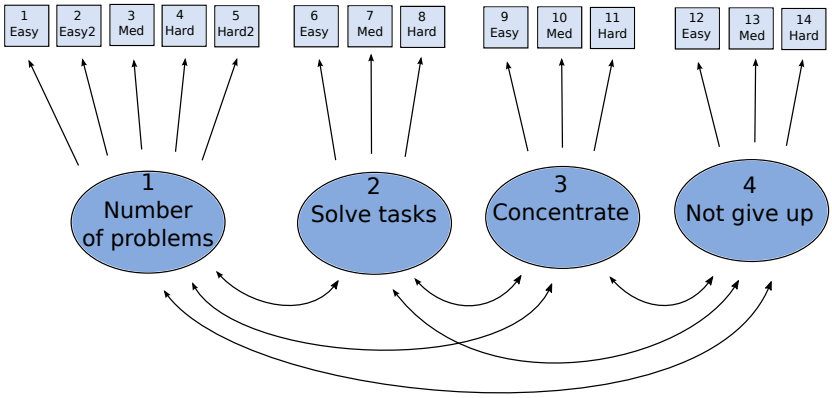


Figure 4: 4 correlated latent facets (Model 2)

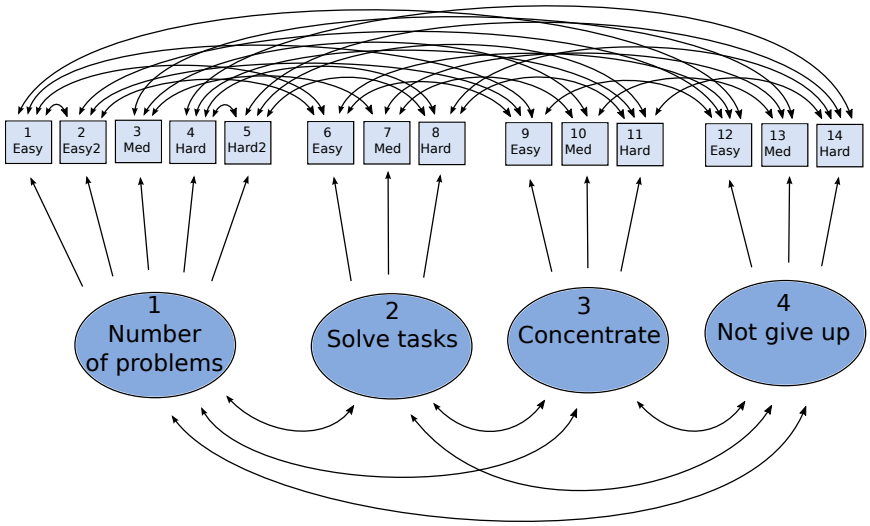


Figure 5: 4 correlated latent facets, with CUs* for 3 levels (Model 2b)

*CUs = correlated uniquenesses

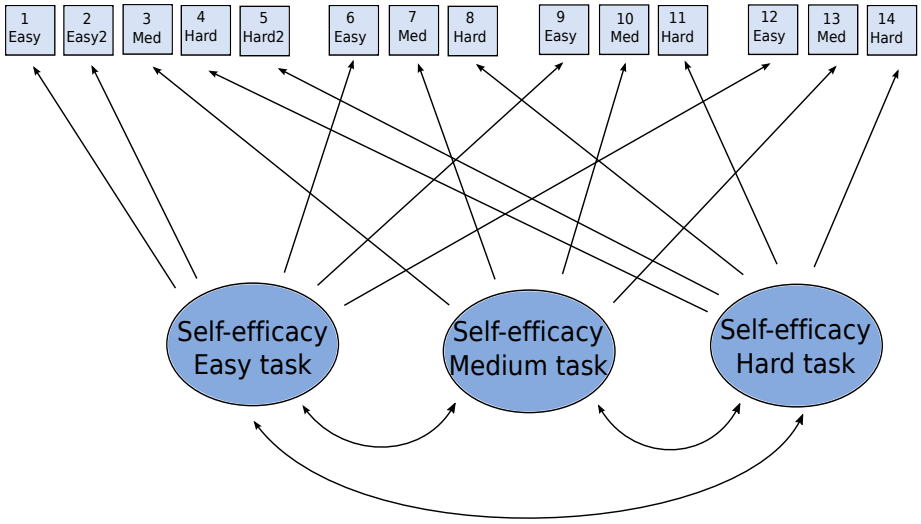


Figure 6: 3 correlated latent levels (Model 3)

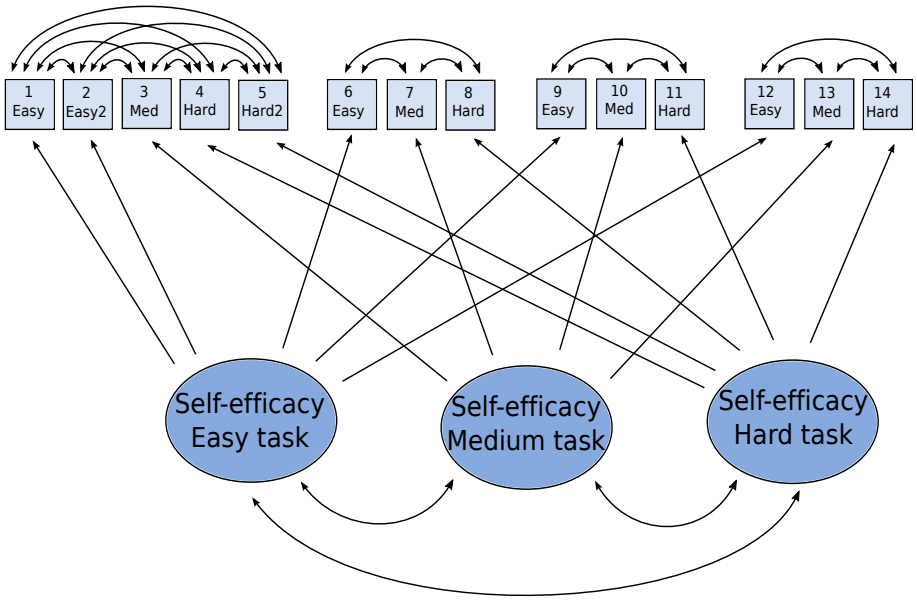


Figure 7: 3 correlated latent levels, with CUs* for 4 facets (Model 3b)

*CUs = correlated uniquenesses

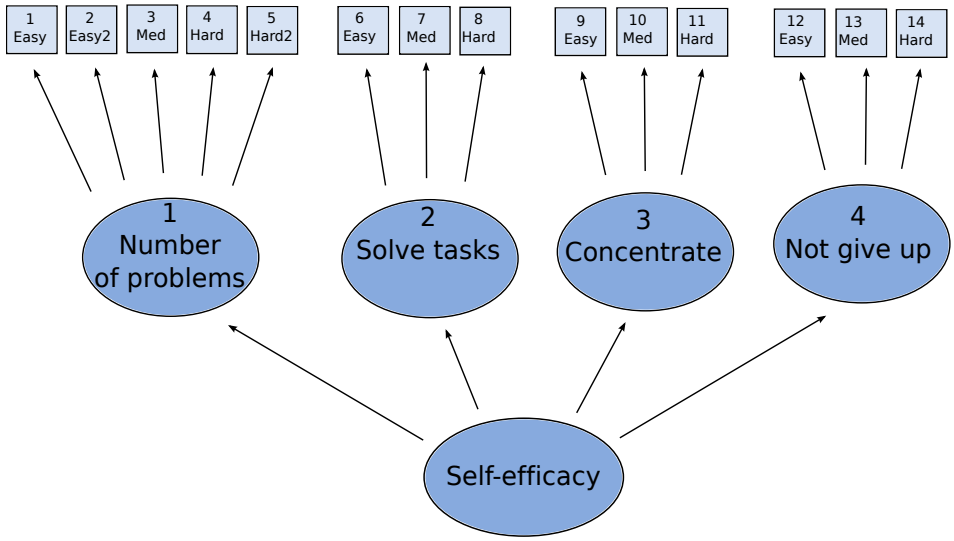


Figure 8: 2nd-order construct with 4 latent facets (Model 4)

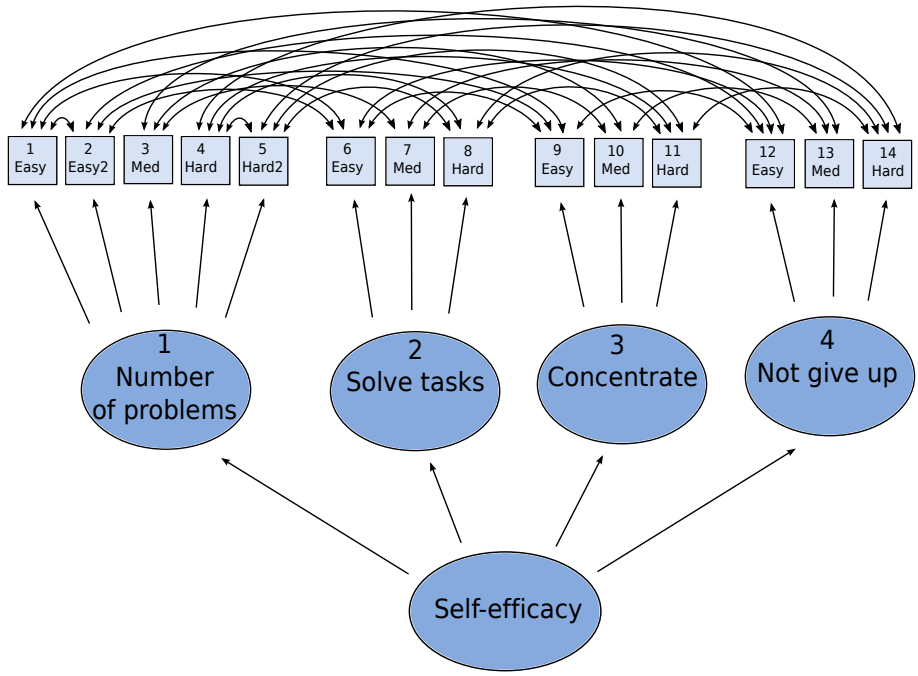


Figure 9: 2nd-order construct with 4 latent facets & CUs* for 3 latent levels (Model 4b)

*CUs = correlated uniquenesses

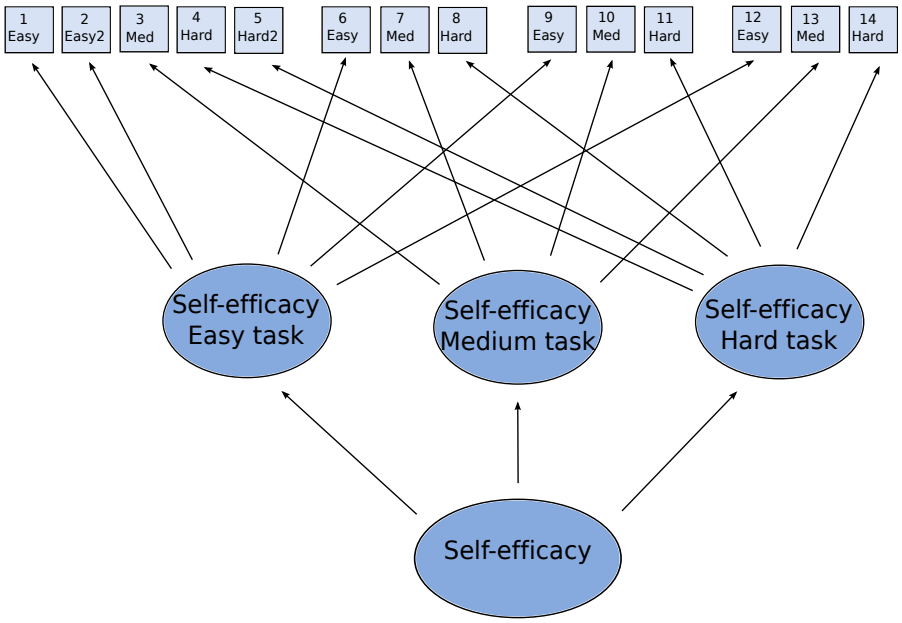


Figure 10: 2nd-order construct with 3 latent levels (Model 5)

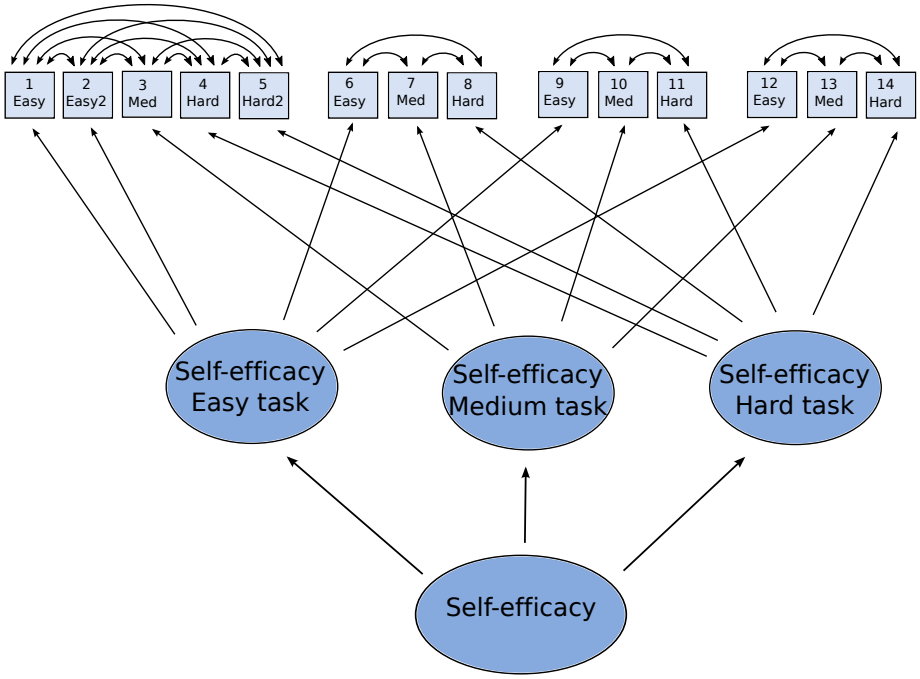


Figure 11: 2nd-order construct with 3 latent levels & CUs* for 4 facets (Model 5b)

*CUs = correlated uniquenesses

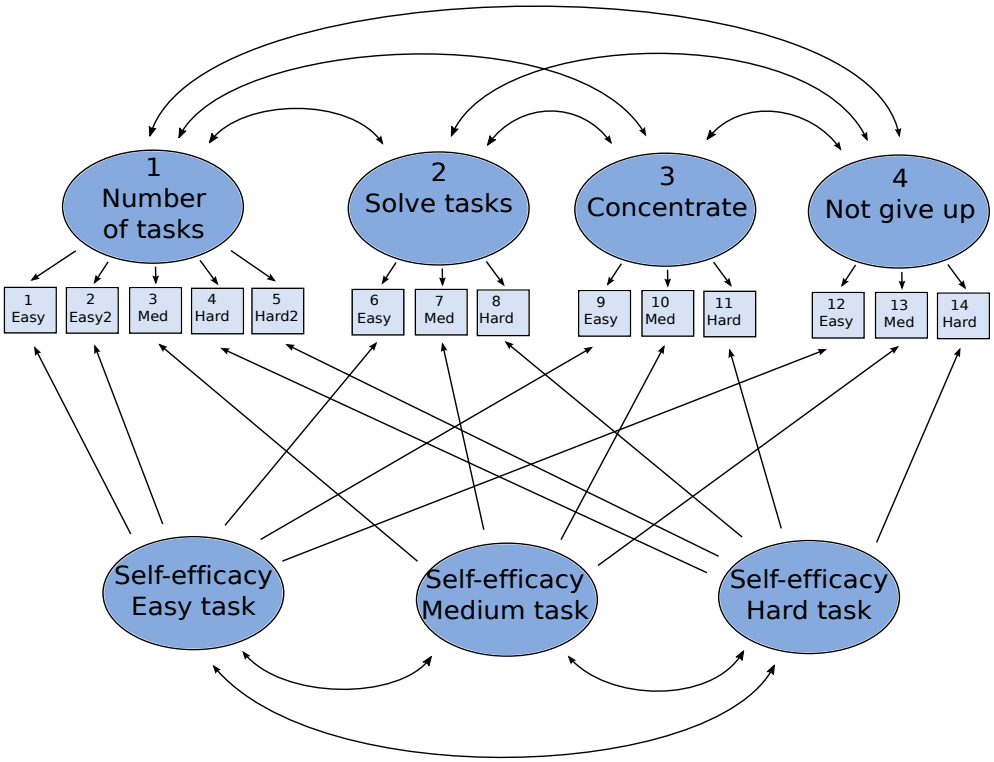


Figure 12: "LMF" model with 4 latent facets & 3 latent levels (Model 6)

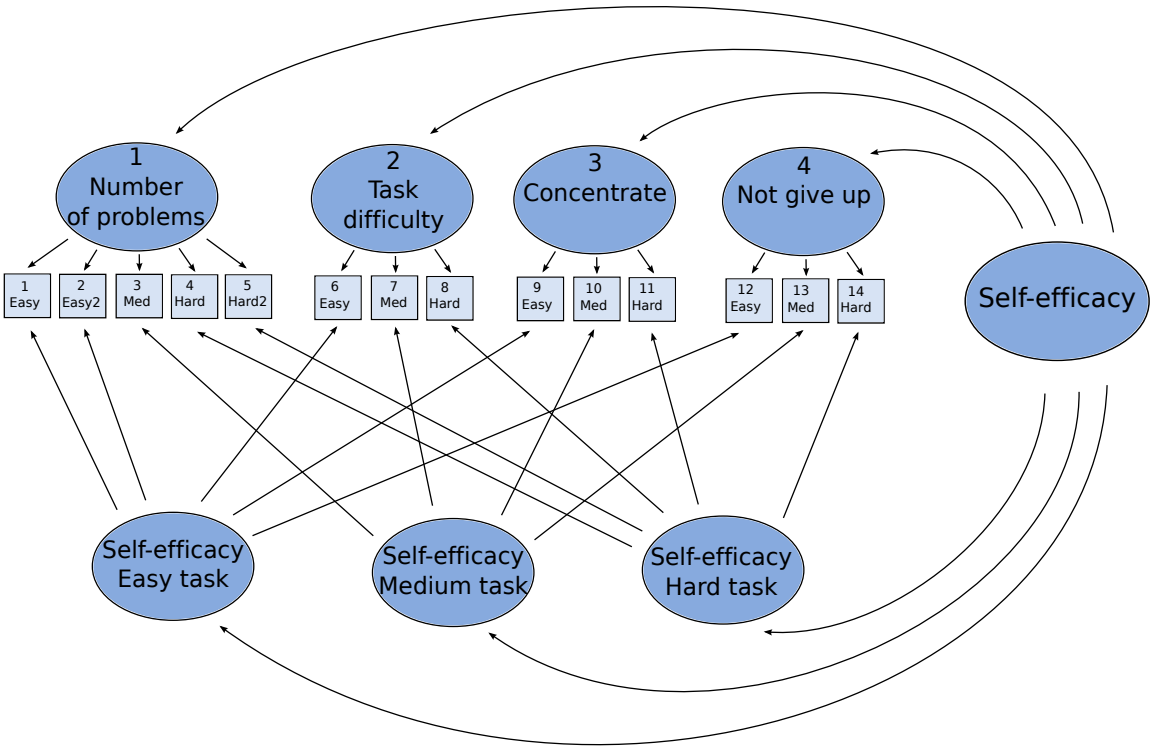


Figure 13: 2nd -order "LMF" model with 4 latent facets & 3 latent levels (Model 7)

2.3.5 Analyses

I used SPSS (version 23) to produce descriptive statistics, while I used the Mplus software (version 1.31 for Mac: Muthén & Muthén, 2012) to run structural equation models. As some items were non-normally distributed I used the robust maximum likelihood (MLR) estimator in Mplus. This estimator uses all available data in accordance with the full-information maximum likelihood (FIML) approach to missingness (Graham & Coffman, 2012). This is in accordance with missing at random (MAR) assumptions. When producing correlation matrices, I used pairwise deletion in SPSS. Data analyses were carried out separately for grade 5 and grades 8/9. For research question one, I first inspected the correlation matrices (Table 8), and then carried out confirmatory factor analyses. For research question two, I used the findings from the confirmatory factor analyses to specify full structural equation models. For research questions three, I used the findings from research question two to specify full structural equation models, including the scores on easy, medium, and hard level national test

items as dependent variables. For research question three I also used descriptives from SPSS to calculate mean accuracy and bias, in order to enable comparison with a previous research study including level-specific self-efficacy in mathematics (Chen and Zimmerman, 2007).

Fit indices

Several fit indices are produced in Mplus to determine model fits, and different researchers have preferred different fit indices. Based on recommendations from researchers in the field (Chiorri, 2014; Morin, Marsh, Nagengast, & Scalas, 2014; Schermelleh-Engel, Moosbrugger, & Müller, 2003) the following fit indices were considered: the χ2 (inferential evaluation of fit), the Root Mean Square Error of Approximation (RMSEA), the Standardized Root Mean Square Residual (SRMR), the Comparative Fit Index (CFI) and the Tucker-Lewis index (TLI) (descriptive evaluations of fit). While there are no absolute guidelines for what values constitute adequate fit, Schermelleh-Engel and colleagues (2003) produced an overview over commonly accepted values for a range of fit indices, including the χ2 (p>.05, or χ2/df =<3 acceptable, <2 good), the RMSEA (<.08 acceptable, <.05 good) the SRMR (<.10 acceptable, <.05 good) and the CFI (> .95 acceptable, >.97 good). Other researchers have used slightly different values to estimate fit, for instance Chiorri (2014): RMSEA (<.06 reasonable), CFI/TLI (>.90 acceptable, >.95 excellent), Hu & Bentler (1999): RMSEA (close to .06), SRMR (close to .08), CFI/TLI (close to .95) and Morin & colleagues (2014): RMSEA (<.08 adequate, <.06 excellent), CFI/TLI (>.90 adequate, >.95 excellent). The χ^2 test statistic is rarely non-significant with large sample sizes, thus several researchers advocate not placing too much emphasis on this, but rather looking at the comparative χ^2 estimate between nested models.

Given the complexity of the models tested (see section 2.3.4) the most lenient of the above-cited fit indices were used to select models with "acceptable" fit for further comparison and substantive consideration. Specifically, as indices of good model fit I used the chi square (χ^2/df =<3 acceptable), the Root Mean Square Error of Approximation (RMSEA <.08 acceptable), the Standardized Root Mean Square Residual (SRMR<.10 acceptable), the Comparative Fit Index (CFI) and the Tucker-Lewis index (TLI) (CFI/TLI > .90 acceptable). Furthermore, models with "acceptable" values on a minimum of three of the five indices were considered.

While the above fit values are related to whether the hypothesized model fit the observed data reasonably well, another issue is relative improvement in fit between competing models. For nested models, the difference in χ^2 in relation to the difference in degrees of freedom can be considered. Furthermore, Chen (2007) suggested the following cut-offs to reject a more parsimonious model (i.e. how much worse the fit of the most parsimonious model can be before it should be rejected): ΔRMSEA (.015), ΔCFI (-.010), ΔSRMR (.030). In my study, after identifying models with acceptable fit, these were compared in regard to improvement in fit relative to model parsimony, using the cut-offs suggested by Chen (2007). Thus, for instance, an alternative model would need to have a CFI value representing at least .010 improvement to that of a more parsimonious model, in order to be retained.

Fit indices provide useful guidelines toward selecting good-fitting models, however several researchers have cautioned that the use of such estimates should not replace sound theoretical judgment (Bollen & Long, 1983; Chiorri, 2014; Little, 1997; Marsh, Dowson, Pietsch, & Walker, 2004; Schermelleh-Engel et al., 2003). Chiorri (2014) as well as Schermelleh-Engel (2003) argue substantive and practical considerations need to be considered when determining

the best model fit, and Marsh & colleagues (Marsh et al., 2004) caution against "logical positivism", arguing that the sizes of standard errors as well as appropriateness of interpretations ought to be considered. It makes little sense to retain a model that gives parameter estimates that are a-theoretical or illogical. An example is Chiorri's (2014) test of the factor structure of the Toronto Empathy Questionnaire (TEQ). While two models specifying correlated uniquenesses (CUs) provided the best model fit, the author argued for retaining an alternative model specifying latent method factors (LMFs). This model gave a slightly worse fit (however with good fit indices), but was considered more practical. Along the same lines Schermelleh-Engel & colleagues (2003) advocated that researchers should consider whether the parameter estimates are meaningful with reasonably sized standard errors, when reviewing model fit. In my study, the retained models were compared in regard to whether the parameter estimates were meaningful with reasonable standard errors, to arrive at a best-fit model.

Dealing with negative residual variances

The occurrence of non-positive residual variances (variably referred to as 'improper solutions' or 'Heywood cases') have challenged researchers since the early days of factor analyses, and the reason for these types of improper solutions in SEM is still not fully understood (Chen, Bollen, Paxton, Curran, & Kirby, 2001). According to some studies as many as 15% of properly specified models might result in improper solutions (Chen et al., 2001) and non-positive residuals are especially likely in the case of multitrait-multimethod models (MTMM) (Chiorri, 2014).

According to Chen & colleagues (2001), researchers should be cautious to use negative error variance estimates as indicators of model misspecification, as the relationship between model

specification and negative error variance is complex, and there can be several reasons for the occurrence of Heywood cases. Some are nonconvergence, underidentification, empirical underidentification, structural misspecification or sampling fluctuations (Kolenikov & Bollen, 2012). Dillon, Kumar and Mulani (1987) suggested model overfitting (i.e. overfactoring) as a likely reason when good model fit was obtained, but with a significant negative residual. Typical approaches to dealing with non-positive residual variances have been to fix the "offending" estimate at zero or at some arbitrarily small positive number such as .005 (Dillon et al., 1987; Gerbing & Anderson, 1987; Morin et al., 2014). However, researchers are advised to investigate the likely cause of non-positive residuals in order to determine how to deal with them (Chen et al., 2001; Kolenikov & Bollen, 2012). If sampling fluctuations are the likely cause then fixing the estimate might be reasonable, while respecification might be necessary if structural misspecification is the cause. Chen and colleagues (2001, p. 503) presented step-by-step recommendations to determine whether negative residuals were likely caused by sampling fluctuations or model misspecification. Several researchers report that it is unusual to come across models with more than one negative estimate (Gerbing & Anderson, 1987; Kolenikov & Bollen, 2012), and no recommendations have been found in regard to how to deal with such cases. When I encountered negative residuals, step-by-step recommendations by Chen et al. (2001) were followed. Models with more than one "Heywood case" were taken as a sign of model misspecification, and rejected.

Calculating accuracy and bias

To explore the relationship between students' self-efficacy and their scores on easy, medium, and hard tasks, I first calculated variables for students' bias and accuracy. I used descriptives derived from SPSS to calculate these variables, in accordance to the methodology adopted by Chen and Zimmerman (2007), to enable comparison between the scores. Accuracy represents

the magnitude of judgment errors of the self-efficacy scores, while bias represents the direction of the judgment errors, in relation to students' scores on the national test items. To calculate bias, students' scores on each item of the national tests (originally 0=wrong, 1=right) were first recoded as 1=wrong and 11=right, to reflect the 1-11 scale of students' self-efficacy strength (Chen and Zimmerman used a 1-8 scale for self-efficacy). The items were rank-ordered according to the percentage of students in my sample that got each item correctly. They were then split into three groups (see Tables 4-7, section 2.3.3), where the test items were placed in the easy (high percentage of students got the item right), medium, or hard group (low percentage of students got the item right). I used the same cut-offs as Chen and Zimmerman (2007) to group the rank-ordered items (difficulty value of at least 65% = easy category, difficulty value of below 50% = hard category). I did not get access to the difficulty value for the first 10 items of the test for 10 students in grade 5, 2013. This is because these students sat a specific test, distributed to a small percentage of the students in the national sample every year, and the test items were hence not released. Although I have the item scores for these students, I do not know which items on their test correspond with which test items in the other test sets. These 10 grade 5 students were thus excluded from the analyses of accuracy and bias.

An obvious advantage to using previously adopted cut-off points is comparability of findings, perhaps especially pertinent given that I have found only few studies that investigate level of self-efficacy in mathematics. Another advantage is that the cut-offs are equivalent for the different grade levels in my sample, while this would not have been the case had I for instance chosen to split the groups to achieve equal number of items in each group. On the other hand, a limitation of this is that the percentage scores (difficulty values) represent difficulty in regard to student mean performance, while there are likely to be individual differences in which items are difficult for which students. This limitation applies to both my own study,

and that by Chen and Zimmerman (2007).

Bias was estimated through deducting the mean of students' national test scores for easy, medium, or hard items, from the mean of their self-efficacy for the corresponding level of difficulty, with a range of scores from -10 (maximum underconfidence) to +10 (maximum overconfidence). Accuracy was calculated through for each student deducting the absolute bias score from 10, with a range of scores from 0 (=complete inaccuracy) to 10 (=complete accuracy).

While my methodology for calculating bias and accuracy is based on the approach taken by Chen and Zimmerman (2007), it differs in that the previous authors matched self-efficacy for specific items with students' scores on similar items, while I match the mean of self-efficacy for a level of perceived difficulty (e.g. self-efficacy for easy level tasks) with the mean of a group of test items (e.g. easy level items). While there would have been benefits to matching self-efficacy and test scores at the item level in my study also, it was in this case not possible, as the test items in my study were from a national test, and thus not available for these purposes.

2.4 Results

2.4.1 Correlation coefficients

In Table 8 I provide correlation coefficients between the 14 self-efficacy items and the overall national test scores. The correlation coefficients for grade 5 and grades 8/9 were inspected separately, to look for patterns in the data. Some overall tendencies emerged across both samples, however. Consistently, each item is most strongly correlated with the item closest to it in regard to level, within the same facet. For example, item 9 (concentrate_easy) is strongly correlated with item 10 (concentrate_med), but less so with item 11 (concentrate_hard). Thus,

items are not always strongly correlated with other items within the same facet, if they are at opposite ends in regard to their respective associations with levels of difficulty. In addition, there is a tendency for strong item correlations between levels of difficulty across facets. An example is item 8 (solve tasks_hard): within each facet it is most strongly correlated with the item(s) associated with the hard level, followed by items associated with the medium and the easy levels. Of the self-efficacy items, five cross-correlations are not significant in the grade 5 score (item 1 with items 4 and 5; item 12 with items 3, 4, and 5), and two in the grade 8/9 scores (item 1 with item 5; item 5 with item 12). The non-significant cross-correlations are consistently associated with easy level items (no. of problems_easy and concentrate_easy), in relation to medium or hard level items. Taken as a whole, these patterns in item correlations indicate relationships both within and across the hypothesized facets and levels, warranting a MTMM type of models. Both levels and facets seem to influence students' responses, which supports the multidimensional models (Models 2b, 3b, 4b, 5b, 6 and 7) over the unidimensional models (Models 1, 2, 3, 4, and 5).

While exploratory factor analysis can give a useful indication of the number of factors and associated factor loading, it is geared towards providing singular factor loadings and thus not well suited to capturing the structure of multidimensional measures. Due to the above findings as well as the theoretical assertion given by Bandura (1997) about the multidimensional nature of self-efficacy, I deemed confirmatory factor analyses as appropriate to investigating the structure of the instrument.

Table 8: Zero-order correlation coefficients among items

	Measure	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	No. of problems_easy		.76**	.45**	.21**	.01	.29**	.31**	.17**	.36**	.24**	.14*	.27**	.24**	.20**	.29**
2	No. of problems_easy2	.61**		.70**	.46**	.20**	.44**	.51**	.35**	.43**	.41**	.34**	.38**	.37**	.30**	.37**
3	No. of problems _med	.32**	.70**		.74**	.41**	.45**	.62**	.51**	.30**	.41**	.44**	.40**	.48**	.42**	.43**
4	No. of problems _hard	.11	0.47**	.75**		.73**	.38**	.60**	.64**	.16**	.33**	.46**	.34**	.42**	.43**	.31**
5	No. of problems_hard2	.09	.33**	.60**	.77**		.25**	.46**	.56**	.03	.23**	.43**	.17**	.25**	.30**	.11*
6	Solve tasks_easy	.26*	.28**	.30**	.28**	.26**		.59**	.35**	.34**	.41**	.36**	.34**	.30**	.21**	.18**
7	Solve tasks_med	.24**	.40**	.53**	.56**	.48**	.37**		.76**	.30**	.42**	.50**	.39**	.49**	.45**	.36**
8	Solve tasks_hard	.15**	.35**	.51**	.56**	.56**	.17*	.72**		.15*	.30**	.50**	.29**	.42**	.49**	.29**
9	Concentrate _easy	.35**	.21*	.09	-.03	-.01	.21*	.16*	.15**		.67**	.24**	.44**	.37**	.22**	.34**
10	Concentrate _med	.24**	.36**	.38**	.31**	.24**	.22**	.48**	.44**	.56**		.64**	.47**	.49**	.39**	.31**
11	Concentrate _hard	.18**	.37**	.43**	.38**	.35**	.20**	.46**	.47**	.26**	.68**		.35**	.48**	.51**	.18**
12	Not give up _easy	.20*	.25**	.25**	.17**	.11*	.31**	.28**	.24**	.35**	.33**	.31**		.80**	.47**	.42**
13	Not give up _med	.28**	.38**	.46**	.39**	.29**	.27**	.44**	.43**	.28**	.43**	.44**	.71**		.79**	.44**
14	Not give up _hard	.23**	.37**	.50**	.42**	.35**	.15**	.45**	.48**	.23**	.40**	.45**	.50**	.87**		.35**
15	National test score	.11**	.27**	.32**	.35**	.28**	.17**	.38**	.38**	.12*	.17**	.26**	.27**	.32**	.25**	

Note: Grade 5 estimates are shown below the diagonal, grade 8/9 estimates are shown above the diagonal.

*Correlation is significant at the p<.05 level (2-tailed)

**Correlation is significant at the p<.01 level (2-tailed)

Estimates are from Mplus (Muthén and Muthén 2012)

In Tables 9, 10, and 11 I present the estimated correlation coefficients for the latent variables, that is, the latent self-efficacy level constructs, and the national test parcels for easy, medium, and hard tasks. These estimates were considered in relation to research question 3, where the relationship between the latent self-efficacy constructs and their 'matched' national test parcels were investigated. The two grade 5 samples are combined, while the grade 8 and grade 9 samples are presented separately. The reason for this is that the number of items in each national test parcel differ between grades 8 and 9 (see section 2.4.3), and there are theoretical reasons the accuracy and bias might differ according to grade level (see section 2.2.2).

Table 9: Zero-order correlation coefficients among latent variables, grade 5

	Measure	1	2	3	4	5	6
1	Self-efficacy, easy level						
2	Self-efficacy, medium level	.91					
3	Self-efficacy, hard level	.84	.99				
4	Test scores, easy items	.46	.42	.39			
5	Test scores, medium items	.31	.34	.33	.72		
6	Test scores, hard items	.32	.37	.38	.67	.73	

Note: All correlations are significant at the p<.01 level (2-tailed)
Estimates are from Mplus (Muthén and Muthén 2012)

In the grade 5 correlation matrix, we see that the highest correlations between self-efficacy and national test parcels are in relation to the easy national test parcels. Interestingly, both medium and hard level self-efficacy is more strongly correlated with easy test items, followed by hard and then medium level test items. Furthermore, we see that each of the three national test parcels are most strongly correlated with the corresponding level of self-efficacy (e.g. easy national test items with self-efficacy for easy tasks). Overall, the estimates related to the medium and hard level self-efficacy constructs are quite similar, particularly in relation to medium and hard national test parcels.

Table 10: Zero-order correlation coefficients among latent variables, grade 8

	Measure	1	2	3	4	5	6
1	Self-efficacy, easy level						
2	Self-efficacy, medium level	.90					
3	Self-efficacy, hard level	.71	.91				
4	Test scores, easy items	.31	.28	.22			
5	Test scores, medium items	.31	.35	.32	.77		
6	Test scores, hard items	.20	.25	.28	.72	.75	

Note: All correlations are significant at the p<.01 level (2-tailed)
Estimates are from Mplus (Muthén and Muthén 2012)

The grade 8 correlation matrix mirrors the grade 5 in that each of the three national test parcels are more strongly correlated with the corresponding self-efficacy construct. Different to the grade 5 matrix, the strongest correlations are in relation to the medium national test parcel. Thus, for each of the self-efficacy level constructs, the strongest correlation is with the medium national test parcel. For the grade 8 sample, the correlation estimates all differ for the different latent variables.

Table 11: Zero-order correlation coefficients among latent variables, grade 9

	Measure	1	2	3	4	5	6
1	Self-efficacy, easy level						
2	Self-efficacy, medium level	.88					
3	Self-efficacy, hard level	.66	.89				
4	Test scores, easy items	.47	.41	.31			
5	Test scores, medium items	.41	.36	.27	.75		
6	Test scores, hard items	.47	.41	.31	.72	.74	

Note: All correlations are significant at the p<.01 level (2-tailed)
Estimates are from Mplus (Muthén and Muthén 2012)

The grade 9 correlation matrix mirrors that of the grade 5 sample, in that the strongest correlations are in association with the easy national test parcel, while the weakest correlations are in association with the medium national test parcel. While the estimates are consistently highest in relation to the easy national test parcel, they are very similar to the estimates associated with the hard national test parcel. Different to both the grade 5 and the grade 8 matrix, all three national test parcels are most strongly correlated with the easy level self-efficacy construct, rather than the medium and hard test

parcels being most strongly correlated with the medium and hard self-efficacy constructs, respectively. The second strongest estimates are in relation to the medium level self-efficacy construct.

2.4.2 RQ1: Confirmatory factor analyses

Confirmatory factor analyses were run testing the empirical models presented in section 2.3.4. Results from the grade 5 and grade 8/9 samples are considered separately below.

Grade 5 results, confirmatory factor analyses

Results from grade 5 confirmatory factor analyses are presented in Table 12. Five models were initially accepted, based on fit indices (see section 1.4.5). Models 2b, 3b, 4b, 5b and 6 were deemed "acceptable" in regard to all fit indices considered. The baseline "unidimensional" model (Model 1) had very poor fit, as did Models 2, 3, 4, and 5. Models 2, 3, 4, and 5 estimate fewer parameters than the other models, which contributes to their worse fit. Model 7 did not converge. The best-fitting models in absolute terms are models 3b and 5b (identical fit values), against which I compared the other models.

In the grade 5 scores, Models 3b and 5b are substantially better fitting than Model 6, based on four out of five Δ fit indices considered (see Table 13). The differences in fit values between Models 3b/ 5b and Models 2b or 4b, respectively, are generally smaller than the recommended cut-offs (ΔRMSEA<.015, ΔCFI<.010, ΔSRMR<.030, $\Delta\chi^2_{(\Delta df=10)}$<18.31 Model 2b, $\Delta\chi^2_{(\Delta df=8)}$<15.51 Model 4b), except CFI (ΔCFI=.010 Model 4b) and TLI (ΔTLI>.010 both models). However, Models 2b and 4b are less parsimonious than Models 3b / 5b. Given that Models 3b and 5b are more parsimonious and have better fit indices, Models 3b and 5b should be preferred.

Table 12: Results from confirmatory factor analyses, grade 5

No:	Model description:	χ^2	p	RMSEA	CFI	TLI	SRMR
1	Unidimensional	$\chi^2_{(77)}$=882.49 χ^2/df=11.46	p<.001	.176	.541	.458	.116
2	Correlated latent facets*	$\chi^2_{(71)}$= 407.28 χ^2/df=5.74	p<.001	.119	.808	.754	.100
2b	**Correlated latent facets with CUs for 3 levels****	$\chi^2_{(45)}$= 121.23 χ^2/df=2.69	p<.001	.071	.957	.912	.069
3	Correlated latent levels	$\chi^2_{(74)}$= 810.69 χ^2/df=10.96	p<.001	.172	.580	.484	.109
3b	**Correlated latent levels with CUs for 4 facets**	$\chi^2_{(55)}$= 115.18 χ^2/df=2.09	p<.001	.057	.966	.943	.057
4	2nd-order latent facets**	$\chi^2_{(73)}$= 416.73 χ^2/df=5.71	p<.001	.118	.804	.756	.104
4b	**2nd-order latent facets with CUs for 3 levels****	$\chi^2_{(47)}$= 124.88 χ^2/df=2.66	p<.001	.070	.956	.914	.073
5	2nd-order latent levels*	$\chi^2_{(74)}$= 810.68 χ^2/df=10.96	p<.001	.172	.580	.484	.109
5b	**2nd-order latent levels with CUs for 4 facets*****	$\chi^2_{(55)}$= 115.18 χ^2/df=2.09	p<.001	.057	.966	.943	.057
6	**Latent constructs model****	$\chi^2_{(54)}$= 158.62 χ^2/df=2.94	p<.001	.076	.940	.900	.069
7	2nd-order latent method factor model by facets & levels	Did not converge					

Models in **bold** type were deemed to have "acceptable" fit indices
*Negative residual variance on 1 item
**Negative residual variance on 2 or more items
***Negative residual variance on 1 factor

Table 13: Δ fit indices grade 5 models*

No:	Model:	$\Delta\chi^2$	Δdf	ΔRMSEA	ΔCFI	ΔTLI	ΔSRMR
2b	Correlated latent facets with CUs for 3 levels	6.05	-10	.014	-.009	-.031	.012
4b	2nd-order latent facets with CUs for 3 levels	9.71	-8	.013	-.010	-.029	.016
6	Latent constructs model	43.45	-1	.019	-.026	-.043	.012

*models were compared against Models 3b /5b

Beside fit indices I considered model parameters, in regard to standard errors, residual variances and whether the estimates were meaningful in substantive terms. Summaries for each of the five initially accepted models are given below.

Correlated facets /correlated levels models with CUs (Models 2b and 3b)

Model 2b results in meaningful and logical estimates in relation to all four facet factors (small standard errors and significant factor loadings). There are significant correlations for all CUs

specified according to the hard level, but some non-significant correlations for CUs according the easy level, in association with items 1 (No. of problems_easy) and 2 (No. of problems_easy2). There are negative residual variances on all four items associated with the medium level (as well as non-significant correlations for CUs specified to the medium level). Correlations between the four facets are all significant and relatively high (smallest is r=.391 between facet 1 "number of problems" and facet 3 "concentrate").

Model 3b results in overall meaningful parameter estimates and small standard errors (the largest standard errors are in relation to the easy level factor, with S.E.>.10 on three items). There are high (>.45 STDYX standardization) and significant loadings on all three level factors. CUs between facet items indicate a clear pattern: There are generally high and significant correlations with the nearest item in the same facet, while correlations are lower and some times non-significant with items further away, within the same facet. The exception is the correlation between items 9 (Concentrate_easy) and 10 (Concentrate_med), which is low and non-significant. There are no negative residual variances. The correlation between the medium and hard level factor is very high (>.97).

Due to the high intercorrelations between the medium and hard level factors in Model 3b, a model was specified similar to Model 3b but where the medium and hard level factors were combined (Model 3c). Model 3c resulted in slightly worse but comparable model results to those of Model 3b. Given that a) this model was not a-priori specified, and b) fit indices did not improve, Model 3c was not retained.

Second-order models with correlated uniquenesses (CUs) (Models 4b and 5b)

Model 4b results in meaningful and significant item loadings on all four facet factors, and factor loadings on the higher order construct self-efficacy, as well as small standard errors. Like with

Model 2b, there are several non-significant correlations associated with items 1 (No. of problems_easy) and 2 (No. of problems_easy2) (CUs according to the easy level factor), and negative residual variances on all items associated with the medium level factor. Correlations specified according to the hard level factor are all significant.

The results from Model 5b are largely identical to those from Model 3b, with overall meaningful parameter estimates (high loadings) and small standard errors on all three level factors, as well as on the factor loadings on the higher order factor self-efficacy (medium factor highest loading). The facet correlations indicate the same pattern as for Model 3b; that correlations are high and significant with the nearest item in the same facet, with the same exception to the pattern; the non-significant correlation between items items 9 (Concentrate_easy) and 10 (Concentrate_med). The model results in negative residual variance on the medium level factor, and was thus subjected to tests recommended by Chen & colleagues (2001) (see section 2.3.5). These tests indicated the negative estimate was unlikely to be a result of sampling fluctuations, and should thus be re-specified (here: rejected).

Latent constructs / method factor (LMF) model (Model 6)

Model 6 results in meaningful and significant loadings (with small standard errors) on all four facet factors, except the loading of item 1 (No. of problems_easy) on facet 1 "number of problems", which is non-significant. Loadings on the easy level factor are all non-significant, where item 1 (No. of problems_easy) is also associated with a very large standard error (S.E.>2). Two loadings on the medium level factor are also non-significant, while factor loadings on the hard level factor are all significant. The facet factors are all meaningfully and significantly intercorrelated, but the easy level factor is not significantly correlated to either the medium or the hard level factors. The medium and hard level factors share a correlation higher than 1 (r=1.14), which is theoretically not

viable. Finally, there are negative residual variances on items 1 (No. of problems_easy) and 13 (Not give up_med).

In summary, Models 2b, 3b, 4b and 5b all result in largely meaningful parameter estimates, with the exception of some individual items. Model 6 results in non-meaningful factor loadings for all items associated with the easy level factor, as well as non-viable factor correlations between the medium and hard level factors. All models except Model 3b result in one or several negative residual variances. In the case of Models 2b and 4b, all items associated with the medium level factor have negative residual variance estimates, which is associated with non-meaningful estimates for correlated uniquenesses between these items. Negative residual variance on several items is unusual (see section 2.3.5), and was taken as a sign of model misspecification.

Overall, these results indicate the best-fitting models when considering fit values are also the best-fitting when considering parameter estimates (substantive interpretations, standard errors and residual variances). Models 5b and 3b resulted in identical fit indices and both gave substantively meaningful factor loadings, however Model 5b resulted in negative residual variance on the medium level factor, and was rejected. Model 3b was thus retained as the best-fitting model, as it resulted in the best model fit indices, meaningful parameter estimates, and gave no negative residual variances.

Grade 8/9 results, confirmatory factor analyses

The results from the grade 8/9 sample were largely similar to those from the grade 5 sample (see Table 14). Five models were initially accepted. As for grade 5, the baseline model (Model 1) resulted in very poor fit indices. Models 2b, 3b, 4b, 5b and 6 were deemed "acceptable". Again as

for grade 5, Models 3b and 5b (identical fit values) are the best-fitting models in absolute terms, against which the other models were compared.

Table 14: Results from confirmatory factor analyses, grades 8 & 9

No:	Model description:	χ^2	p	RMSEA	CFI	TLI	SRMR
1	Unidimensional	$\chi^2_{(77)}$=1343.40 χ^2/df=17.45	p<.001	.210	.503	.413	.119
2	Correlated latent facets*	$\chi^2_{(71)}$=796.27 χ^2/df=11.21	p<.001	.165	.716	.636	.120
2b	**Correlated latent facets with CUs for 3 levels****	$\chi^2_{(45)}$= 150.40 χ^2/df=3.34	p<.001	.079	.959	.916	.084
3	Correlated latent levels	$\chi^2_{(74)}$=1622.57 χ^2/df=21.93	p<.001	.237	.393	.253	.114
3b	**Correlated latent levels with CUs for 4 facets**	$\chi^2_{(55)}$= 99.07 χ^2/df=1.80	p<.001	.046	.983	.971	.052
4	2nd-order correlated facets **	$\chi^2_{(73)}$=804.65 χ^2/df=11.02	p<.001	.164	.713	.642	.129
4b	**2nd-order correlated facets with CUs for 3 levels****	$\chi^2_{(47)}$= 168.48 χ^2/df=3.58	p<.001	.083	.952	.908	.089
5	2nd-order latent levels***	$\chi^2_{(74)}$=1622.55 χ^2/df=21.93	p<.001	.237	.393	.253	.114
5b	**2nd-order latent levels with CUs for 4 facets *****	$\chi^2_{(55)}$= 99.07 χ^2/df=1.80	p<.001	.046	.983	.971	.052
6	**Latent constructs model****	$\chi^2_{(54)}$= 196.99 χ^2/df=3.65	p<.001	.084	.944	.906	.065
7	2nd-order latent constructs model by facets and levels	Did not converge					

Models in **bold** type were deemed to have "acceptable" fit indices
*Negative residual variance on 1 item
**Negative residual variance on 2 or more items
***Negative residual variance on 1 factor

When considering differences in fit indices (see Table 15), the estimates indicate Models 3b and 5b are substantially better than all the alternative models deemed "acceptable", according to the guidelines currently adopted. The one exception is ΔSRMR for Model 6, which is lower than the recommended cut-off of .030.

Table 15: Δ fit indices grade 8 & 9 models*

No:	Model description:	$\Delta\chi^2$	Δdf	ΔRMSEA	ΔCFI	ΔTLI	ΔSRMR
2b	Correlated dimensions with CUs for 3 levels	51.33	-10	.033	-.024	-.055	.032
4b	2nd-order correlated dimensions with CUs for 3 levels	69.42	-8	.037	-.031	-.063	.037
6	Latent method factor model (CTCM model)	97.92	-1	.038	-.039	-.065	.013

*models were compared against models 3b /5b

Like in the grade 5 sample, Models 3b and 5b are both more parsimonious and have better fit indices, and are thus the preferred models according to fit indices. In the case of grade 8/9 there are bigger differences in fit indices between the best and the second best-fitting models, compared to the grade 5 results.

Model parameters (meaningfulness of the estimates, standard errors and residual variances) were considered for all the initially accepted models for the grade 8/9 data, as was the case for the grade 5 data. A summary for each of the five models is given below.

Correlated facets /correlated levels models with CUs (Models 2b and 3b)

Model 2b results in meaningful and significant estimates for item loadings on all facet factors, significant correlations between facet factors (estimates between .43 and .66), and small standard errors. Correlations between easy and hard level items are all significant except for item 1 (No. of problems_easy) with item 9 (Concentrate_easy). There are negative residual variances on all (four) items associated with the medium level factor, and thus correlations between the medium level items are not estimated.

Model 3b results in meaningful estimates for all item loadings on the level factors (all above .45 and significant), and a meaningful pattern to the estimates of correlation between facet factors. Similar to the grade 5 results, there are generally high and always significant correlations between the nearest items within the same dimension factor, and non-significant correlations between items further away within the same facet (except in the case of facet 4, "not give up", where there are significant correlations between all items). There are small standard errors associated with all estimates in the model (<.09). Unlike the grade 5 results, intercorrelations between easy, medium

and hard level factors are high (>.68), but not above .90. The model results in no negative residual variances.

Second-order models with correlated uniquenesses (CUs) (Models 4b and 5b)

Model 4b results are largely similar to those from Model 2b. Estimates for item loadings on all facet factors are meaningful, and factor loadings on the second order factor are high (>.62) and significant. There are small standard errors. Correlations between the easy and hard level items are all significant except for item 1 (No. of problems_easy) with item 9 (Concentrate_easy). There are negative residual variances on all items associated with the medium level factor, and correlations between medium level items are thus not estimated.

Model 5b, like Model 3b, results in meaningful and significant parameter estimates for item loadings on the level factors. The medium level factor has the highest loading on the second order self-efficacy construct, and factor loadings on the second order construct as well as correlations between facet items are meaningful. There are small standard errors. Like for grade 5, there is negative residual variance on the medium level factor. The recommended tests by Chen & colleagues (2001) indicated this was unlikely due to sampling fluctuations.

Latent constructs / method factor (LMF) model (Model 6)

Model 6 results in meaningful and significant item loadings on all three level factors, and on three out of four facet factors. There are small standard errors associated with the medium and hard level factors, and with the "solve tasks" and "not give up" facets. There is a high standard error (.27) associated with item 10 (Concentrate_med) when loading on facet 3, "concentrate", and relatively high (>.12) standard errors for all items on facet 1, "number of problems", as well as two of these item estimates when loading on the easy level. The "number of problems" facet is also associated

with several problematic item loadings (negative or non-significant). Unlike the grade 5 results, in the case of grades 8/9 Model 6 results in meaningful and significant correlations between the three level factors (between .68 and .90), but low and non-significant correlations between all facet factors. There is negative residual variance on three of four items associated with the medium level factor.

In summary, grade 8/9 results are largely similar to those for grade 5, in that Models 2b, 3b, 4b and 5b all result in largely meaningful parameter estimates, while model 6 results in several substantively problematic estimates. Furthermore, Models 2b and 4b result in negative residual variances on all four items associated with the medium level factor, while Model 5b results in negative residual variance on the medium level factor itself, and was rejected based on recommended tests. All models except from Model 3b result in one or several negative residual variances. As for the grade 5 sample, Model 3b was retained as the best fitting model, based on both fit-indices and consideration of parameter estimates.

Overall, these results indicate that the best-fitting model when considering fit values is also the best-fitting in regard to parameter estimates (substantive interpretations, standard errors, and residual variances). Models 3b and 5b result in identical fit indices, however Model 5b has negative residual variance on the medium level factor, and was rejected. In the case of both grade 5 and grade 8/9 scores the best-fitting model was a correlated levels model with correlated uniquenesses (CUs) specified according to four facets (Model 3b). This model results in the best model fit indices, meaningful parameter estimates, and no negative residual variances. In Model 3b, latent correlations are .82 (for grade 5) and .89 (for grades 8/9) between the easy and medium levels, .97 (for grade 5) and .90 (for grades 8/9) between the medium and hard levels, and .68 (for both grade 5 and grades 8/9) between the easy and hard levels.

The theoretical model proposed in section 2.2.4 was tested with empirical Model 7, a second order model including latent facet and latent level factors. This model did not converge for either of the samples.

2.4.3 RQ2: Associations between national test results and self-efficacy

Having selected a best-fitting measurement model (Model 3b), I used this as the measurement part of a full structural equation model including national test scores as the correlate. To avoid issues of co-linearity due to the high correlation between the medium and hard level constructs (see section 2.4.2), national test scores were correlated with, rather than regressed on, the three level constructs. For comparison purposes I also specified as full structural equation models the unidimensional model (Model 1), substantively meaningful measurement models with "acceptable" model fit (Models 2b, 4b, and 5b), and my theoretical model (Model 7). Model 7 does not converge as measurement model alone, but when test scores are included it does. Results from the structural equation models are presented separately for the grade 5 and grade 8/9 samples below.

Grade 5 results: Structural equation models

Results from the grade 5 sample are presented in Table 16. As expected, the unidimensional model results in poor model fit, while Models 2b, 3b, 4b, and 5b all result in "acceptable" fit indices. When specified as full structural equation model, the theoretical model (Model 7) converges, and results in acceptable fit indices. The best-fitting model in absolute terms is Model 3b.

Table 16: Results from structural equation models, grade 5

No:	Model description:	χ^2	p	RMSEA	CFI	TLI	SRMR
1	Test scores correlated w/ unidimensional model	$\chi^2_{(90)}$= 946.27 χ^2/df=10.51	p<.001	.162	.548	.473	.110
2b	Test scores correlated w/ latent facets w/ CUs for levels**	$\chi^2_{(55)}$=159.58 χ^2/df=2.90	p<.001	.072	.945	.895	.075
3b	Test scores correlated w/ latent levels w/ CUs for facets	$\chi^2_{(66)}$= 140.80 χ^2/df=2.13	p<.001	.056	.961	.937	.056
4b	Test scores correlated w/ 2nd-order latent facets w/ CUs for levels**	$\chi^2_{(60)}$=172.40 χ^2/df=2.87	p<.001	.072	.941	.896	.076
5b	Test scores correlated w/ 2nd-order latent levels w/ CUs for facets***	$\chi^2_{(68)}$= 147.38 χ^2/df=2.17	p<.001	.057	.958	.935	.056
5c	Test scores correlated w/ 2nd-order latent levels w/ CUs for facets, medium factor variance set at .005	$\chi^2_{(69)}$= 150.70 χ^2/df=2.18	p<.001	.057	.957	.934	.058
7	Test scores correlated w/ 2nd order latent constructs model by facets and levels**	$\chi^2_{(69)}$= 174.71 χ^2/df=2.53	p<.001	.065	.944	.915	.061

**Negative residual variance on 2 or more items
***Negative residual variance on 1 factor

When comparing the best-fitting model (Model 3b) to the models with "acceptable" model fit (see Table 17), Model 3b is better fitting than Models 2b, 4b, and 7, while the differences in fit between Models 3b and 5b are below the chosen cut-offs (see section 2.3.5).

Table 17: Δ fit indices grade 5 structural equation models*

No:	Model description:	$\Delta\chi^2$	Δdf	ΔRMSEA	ΔCFI	ΔTLI	ΔSRMR
2b	Test scores correlated w/ latent facets w/ CUs for levels	18.79	-9	.016	-.016	-.042	.019
4b	Test scores correlated w/ 2nd-order latent facets w/ CUs for levels	31.60	-6	.016	-.020	-.041	.020
5b	Test scores correlated w/ 2nd-order latent levels with CUs for facets	6.58	2	.001	-.003	-.002	.000
5c	Test scores correlated w/ 2nd-order latent levels with CUs for facets, medium factor variance set at .005	9.90	3	.001	-.004	-.003	.002
7	Test scores correlated w/ 2nd order latent constructs model by facets and levels	33.91	3	.009	-.017	-.022	.005

*models were compared against Model 3b

Again, I considered model parameters. Models 1 and 3b both result in meaningful factor loadings, small standard errors, and no negative residual variances. In Model 1, the correlation between overall self-efficacy strength and performance scores is .46 for grade 5. Like in the measurement

model, Model 3b results in meaningful parameter estimates for the correlations between test scores and the easy (.43), medium (.47) and hard (.46) level factors. In contrast, Models 2b, 4b, 5b, and 7 all result in one or several negative residual variance estimates:

Model 2b results in negative residual variance on three items associated with the medium level factor. The model also results in several non-theoretical estimates for national test scores correlated with the four facet factors. Of these, only the estimate of national test scores in relation to the "solve task" facet is meaningful (.37) and significant.

Model 4b results in negative residual variances on three items related to the medium level factor (the same three items as Model 2b). However, the estimate for national test scores regressed on the second-order self-efficacy factor (.36) is meaningful and significant.

Model 5b results in one negative residual estimate of variance on the medium level factor, and in associated non-meaningful parameter estimates. The recommended test by Chen and colleagues (2001) indicated the negative residual was due to sampling fluctuations, and thus that it was reasonable to re-specify the model with the variance of the medium level factor set at a non-negative value. Model 5b was thus re-specified as Model 5c, where variance for the medium level factor was set at .005. In this model, parameter estimates for the associations between the higher order self-efficacy construct and the medium and hard level factors are very high (>.98 for both estimates), but viable. Furthermore, the estimate for national test scores correlated with the second-order self-efficacy construct is high (.47) (the same estimate is .45 in model 5b), significant, and comparable to the same estimate from the unidimensional model (Model 1 estimate = .46). Fit indices from Model 5b do not improve when fixing the negative residual to a positive value in Model 5c.

The theoretical model (Model 7) results in non-significant factor loadings and/or large standard errors for most items on three of the seven first order factors, as well as for the level factors on the self-efficacy construct. Also, there are negative residual variances on all items associated with the medium level factor, as well as the medium level construct itself.

Considering both fit indices and parameter estimates, Model 3b is retained as the best-fitting model for the grade 5 sample. Model 3b is the only structural equation model that results in acceptable fit indices and has no negative residual variances. Model 5c results in practically indistinguishable fit-indices to Model 3b, and is also more parsimonious. However, it is specified on the basis of a measurement model with negative residual variance on the medium level factor, where recommended tests indicated it was not reasonable to fix this variance to a positive number. These results are confounding, and it is deemed more consistent to retain Model 3b, which is based on a measurement model without negative residual variances.

Grade 8/9 results: Structural equation models

Results from the grade 8/9 sample are presented in Table 18. Like in the grade 5 results, the unidimensional model (Model 1) results in poor model fit, Models 3b, 5b, and 7 result in "acceptable" fit indices, and the best-fitting model in absolute terms is Model 3b. Unlike the grade 5 sample Models 2b and 4b result in "acceptable" fit indices on less than three estimates each, and are thus not accepted according to the chosen criteria (see section 2.3.5).

Table 18: Results from structural equation models, grades 8 & 9

No:	Model description:	χ^2	p	RMSEA	CFI	TLI	SRMR
1	Test scores correlated w/ unidimensional model	$\chi^2_{(90)}$=1424.55 χ^2/df=15.83	p<.001	.194	.511	.429	.114
2b	Test scores correlated w/ latent facets with CUs for 3 levels**	$\chi^2_{(55)}$=213.33 χ^2/df=3.87	p<.001	.086	.942	.889	.095
3b	Test scores correlated w/ latent levels w/ CUs for facets	$\chi^2_{(66)}$= 139.65 χ^2/df=2.12	p<.001	.053	.973	.957	.056
4b	Test scores correlated w/ 2nd order latent facets with CUs for levels**	$\chi^2_{(60)}$=244.97 χ^2/df=4.08	p<.001	.089	.932	.881	.100
5b	Test scores correlated w/ 2nd order latent levels with CUs for facets***	$\chi^2_{(68)}$= 148.20 χ^2/df=2.18	p<.001	.055	.971	.955	.060
5c	Test scores correlated w/ 2nd order latent levels with CUs for facets, medium factor variance set to .005	$\chi^2_{(69)}$= 168.44 χ^2/df=2.44	p<.001	.061	.964	.945	.062
7	Test scores correlated w/ 2nd order latent constructs model by facets and levels**	$\chi^2_{(69)}$= 226.87 χ^2/df=3.29	p<.001	.076	.942	.912	.062

**Negative residual variance on 2 or more items
***Negative residual variance on 1 factor

The differences in fit between the best-fitting model and the accepted models are presented in Table 19. Like in the grade 5 sample, Model 3b is better fitting than Model 7, and the differences in fit between Model 3b and Model 5b are below the chosen cut-offs.

Table 19: Δ fit indices grades 8 and 9 structural equation models*

No:	Model:	$\Delta\chi^2$	Δdf	ΔRMSEA	ΔCFI	ΔTLI	ΔSRMR
5b	Test scores correlated w/ 2nd order latent levels with CUs for facets***	8.56	2	.002	-.002	-.002	.004
5c	Test scores correlated w/ 2nd order latent levels with CUs for facets, variance set to .005	28.80	3	.008	-.009	-.012	.006
7	Test scores correlated w/ 2nd order latent constructs model by facets and levels	87.22	3	.023	-.031	-.045	.006

*models were compared to Model 3b

Again, parameter estimates were considered in addition to fit indices. Models 1 and 3b both result in meaningful factor loadings, and they are the only models that do not result in any negative residual variances. In Model 1, the correlation between overall self-efficacy strength and performance scores is .48 for grades 8/9. Model 3b results in meaningful and significant estimates of the relationships between level factors and national test scores. The highest estimate is the

association between national test scores and the medium level factor (.51), followed by the easy level factor (.50) and the hard level factor (.38). Models 2b, 4b, 5b and 7 all result in negative residual variances. Of these, only Models 5b and 7 result in otherwise "acceptable" fit indices, and are considered next in regard to parameter estimates.

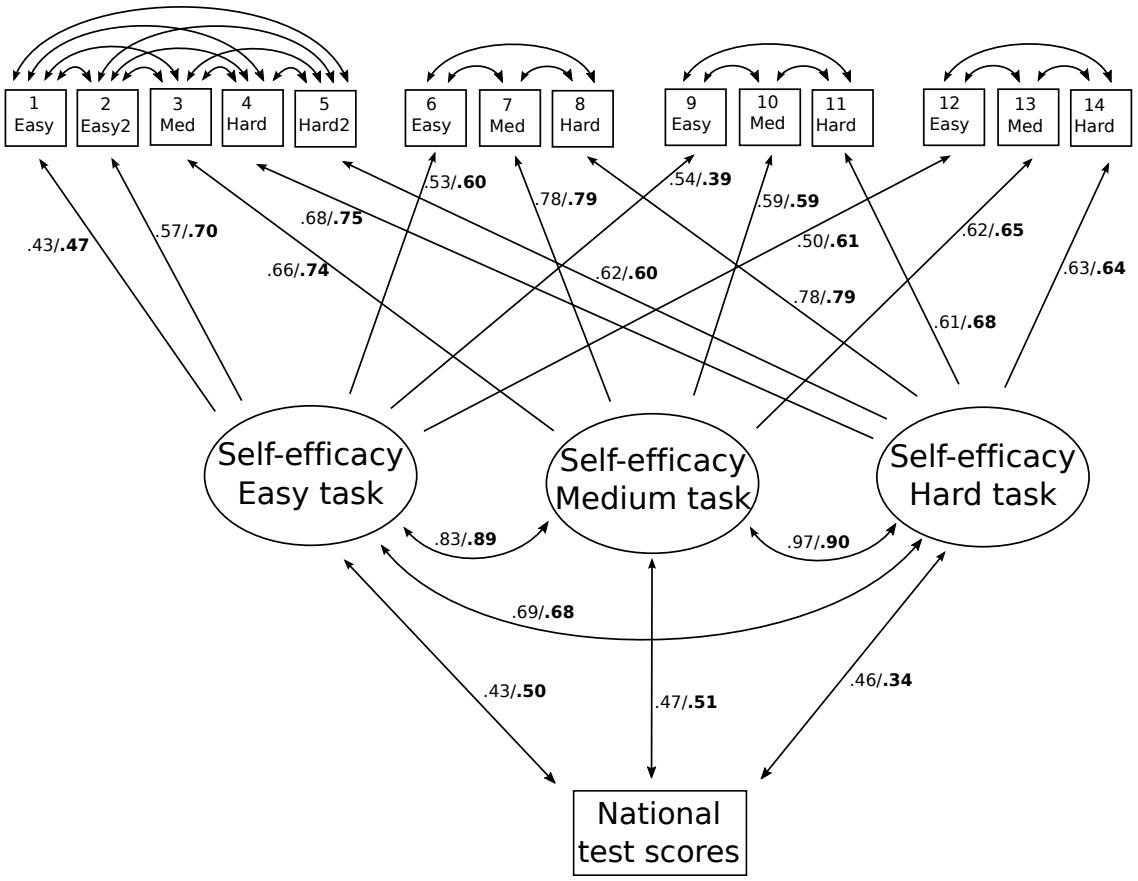
Like for grade 5, Model 5b results in negative residual variance on the medium level factor. Different to the grade 5 sample, the recommended tests indicated the negative estimate was not likely due to sampling fluctuations. Nevertheless, for comparison purposes, Model 5b was respecified as Model 5c, where the variance of the medium level factor was set at .005. This model results in very high estimates for the relationship between all level factors (easy, medium, and hard) and national test scores (>.90 for all three estimates). The relationship between the higher order self-efficacy construct and national test scores is slightly stronger (.49) than that in Model 5b (.45). Setting the variance of the medium level factor to .005 in Model 5c does not lead to improvement in fit indices, as compared to Model 5b. Overall, when I considered fit indices as well as parameter estimates, Model 3b is the best-fitting meaningful model when considering the grade 8/9 sample.

Summary

For both samples, a structural equation model based on the unidimensional measurement model (Model 1) results in poor fit, while Models 3b, 5b, and 7 result in acceptable fit indices. Models 2b and 4b result in "acceptable" fit indices in regard to he grade 5 sample, only. Fit indices for both samples clearly indicate Model 3b as best-fitting, as expected. When considering both fit indices and parameter estimates, Model 3b is retained as the best-fitting structural equation model. In this model the relationship between the latent level factors and national test scores are specified as correlations rather than regression coefficients. The design of the study allows us to interpret these correlation estimates as predictions, given that students completed the self-efficacy questionnaires

prior to taking the national tests. The wording of the questionnaire items further supports this, as items are related to expectations about future performance on the test. It does not make theoretical sense to assume the relationship to be in the opposite direction.

Figure 14 presents factor loadings from Model 3b for grade 5 and grades 8/9. For both samples, students' scores on national tests in mathematics are related to level of their self-efficacy. For all three levels of perceived difficulty, stronger self-efficacy is associated with higher overall national test results. For both samples, this relationship is strongest for national test scores and medium level difficulty tasks.



Note: Estimates before "/" are Grade 5 results, after "/", in **bold**, are grade 8/9 results.
All estimates are standardized (STYX) estimates from Mplus (Muthén & Muthén 2012).
Estimates for CUs are not given due to space limitations.

Figure 14: Structural equation model (Model 3b)

In the case of the grade 5 sample, the relationship with test scores is stronger for the medium (.47) and hard (.46) levels of difficulty, than for the easy level (.43). In the case of grade 8/9 students, the relationship between self-efficacy level and national test scores is stronger for the medium (.51) and easy (.50) levels, than for the hard level (.38) of perceived difficulty. Differences in fit between models are larger in the older student sample.

Applying filters for class response rates

The best-fitting models were tested with different filters: that is, class response rates were used as selection criteria for the analyses. The results as they have been presented up to this point, are in relation to the complete sample (however not including repeat participants in grade 9). In order to investigate whether the findings were influenced by class response rates (see section 2.3.1), I re-specified the best-fitting model (Models 3b) including only students from classes with the highest response rate (above 74%). For the grade 5 and grade 8/9 samples alike, the differences in fit indices were indistinguishable for these models – that is, the differences were smaller than the applied cut-offs (see section 2.3.5). In case of the grade 8/9 sample, the estimates for the associations between national test scores and self-efficacy easy (.49), medium (.50), and hard (.35) levels for the "high response rate only" sample (n= 326) are comparable to the results from the full grade 8/9 sample (see Figure 14). In contrast, the estimates from the grade 5 "high response rate only" sample (n=248) of the associations between national test scores and the easy (.49), medium (.52), and hard (48) levels of self-efficacy differ from the full grade 5 sample (see Figure 14). The most interesting difference is that in the restricted grade 5 sample the estimate is higher for the easy than for the hard level tasks, and thus mirror the results from the grade 8/9 sample, where the strongest estimate is in relation to the medium, followed by the easy, then the hard, level tasks.

2.4.4 RQ3: Self-efficacy and national test scores for easy, medium, and hard tasks

To explore students' self-efficacy relative to their scores on national test items, I calculated the accuracy (the inverse magnitude of students' judgment error, in other words their judgment accuracy) and bias (the direction of students' judgment error) of their self-efficacy for easy, medium, and hard tasks. These estimates were produced to allow comparison with a previous study (Chen and Zimmerman, 2007), which is why I present the scores according to gender, also. Girls' and boys' mean scores of accuracy and bias are presented in Tables 20 through 22, organised by task difficulty (easy, medium, or hard tasks). Mean scores for student self-efficacy and test performance are also presented. My scores range from 1-11, while Chen and Zimmerman (2007) used a 1-8 scale.

Table 20: Accuracy and bias, grade 5

	Easy level		Medium level		Hard level	
	M	SD	M	SD	M	SD
			Grade 5 (n = 363)			
Accuracy	8,33	1,61	7,60	1,92	7,80	1,66
Bias	1,36	1,88	1,58	2,64	0,89	2,62
Self-Efficacy	10,06	1,29	8,42	1,92	5,97	2,32
Test performance	8,64	1,82	6,75	2,50	4,99	2,33
			Grade 5, girls (n= 186)			
Accuracy	8,46	1,41	7,64	1,74	7,77	1,75
Bias	1,24	1,68	1,49	2,53	0,80	2,73
Self-Efficacy	9,87	1,42	8,12	1,98	5,60	2,32
Test performance	8,62	1,66	6,61	2,27	4,73	2,13
			Grade 5, boys (n = 177)			
Accuracy	8,18	1,80	7,56	2,12	7,83	1,56
Bias	1,49	2,08	1,68	2,76	0,98	2,50
Self-Efficacy	10,26	1,11	8,75	1,80	6,37	2,26
Test performance	8,67	1,99	6,90	2,73	5,27	2,50

The grade 5 scores from 2012 and 2013 were investigated separately, and combined, as there seemed to be no systematic differences to the results. The descriptives from Table 20 indicate the grade 5 students in my sample are generally more accurate at judging their capabilities for easy

tasks followed by hard tasks, while they are least accurate in relation to medium tasks. Furthermore, students tend to be positively biased (e.g. to overestimate their capabilities), as the mean bias scores are all in the positive direction. Grade 5 girls in my sample have higher mean accuracy scores than boys in relation to easy and medium, but not hard tasks.

Table 21 indicates some similar tendencies in the grade 8 scores, in that the mean bias scores are positive, indicating a tendency toward overconfidence, as opposed to underconfidence. However, the descriptives indicate the grade 8 students in my sample are progressively worse at judging their capabilities as the tasks get more difficult, demonstrated in how the highest mean accuracy scores are in relation to easy tasks, followed by medium and then hard tasks. Among the grade 8 students in my sample, the girls have consistently higher mean accuracy scores than the boys, in relation to all levels of difficulty.

Table 21: Accuracy and bias, grade 8

	Easy level		Medium level		Hard level	
	M	SD	M	SD	M	SD
			Grade 8 (n= 228)			
Accuracy	7,87	1,67	7,51	1,81	7,35	1,82
Bias	1,94	1,89	2,03	2,31	2,04	2,48
Self-Efficacy	10,15	1,15	8,60	1,71	5,98	2,11
Test performance	8,25	2,07	6,64	2,53	4,03	2,02
			Grade 8, girls (n= 101)			
Accuracy	8,04	1,63	7,63	1,63	7,56	1,76
Bias	1,82	1,78	1,89	2,18	1,74	2,46
Self-Efficacy	10,13	1,23	8,58	1,79	5,89	2,30
Test performance	8,35	2,04	6,67	2,51	4,16	2,01
			Grade 8, boys (n= 127)			
Accuracy	7,71	1,69	7,40	1,95	7,17	1,86
Bias	2,05	1,98	2,16	2,43	2,31	2,48
Self-Efficacy	10,17	1,09	8,61	1,64	6,04	1,95
Test performance	8,20	2,10	6,63	2,58	3,95	2,05

Grade 9 descriptives are presented in Table 22. The grade 9 descriptives share similarities with the grade 5 estimates, in that the students in my sample have higher mean accuracy scores for easy tasks, followed by hard and then medium tasks. Similar to the grade 8 descriptives, the students tend to be positively biased (overconfident), and the girls have higher mean accuracy scores than the boys.

Looking at Tables 20 - 22 overall some major tendencies can be seen across the grade levels. First of all, students across all the sub-samples seem to tend towards overconfidence, rather than underconfidence, displayed in their positive mean bias scores. Second, students in all the grades seem to be more accurate at judging their capabilities in regard to easy level tasks. The estimates differ for grade 5 and grade 9 students, versus grade 8 students, in regard to whether they subsequently have higher mean accuracy scores in relation to medium (grade 8) or hard (grade 5 and grade 9) tasks. In both grades 8 and 9, girls have higher estimates of mean accuracy than boys in relation to all levels of difficulty, while this is true in regard to easy and medium level tasks for grade 5 students. The higher accuracy of the girls relative to the boys is due mainly to their lower bias: while the girls in my sample tend to achieve equally well as the boys (boys have slightly higher test scores in grade 5, while girls have slightly higher test scores in grade 8 and grade 9), they tend to hold slightly lower, and more accurate, self-efficacy.

Table 22: Accuracy and bias, grade 9

	Easy level		Medium level		Hard level	
	M	SD	M	SD	M	SD
			Grade 9 (n= 260)			
Accuracy	8,24	1,56	7,38	1,84	7,78	1,81
Bias	1,56	1,76	2,16	2,36	1,26	2,58
Self-Efficacy	10,18	1,34	8,73	1,89	5,79	2,35
Test performance	8,59	1,92	6,59	2,36	4,54	2,28
			Grade 9, girls (n= 121)			
Accuracy	8,32	1,60	7,50	1,88	7,96	1,64
Bias	1,51	1,77	1,97	2,44	1,00	2,42
Self-Efficacy	10,27	1,31	8,69	1,82	5,65	2,28
Test performance	8,67	1,91	6,67	2,41	4,56	2,34
			Grade 9, boys (n= 139)			
Accuracy	8,17	1,53	7,27	1,80	7,60	1,96
Bias	1,61	1,76	2,35	2,28	1,52	2,70
Self-Efficacy	10,09	1,37	8,77	1,97	5,91	2,42
Test performance	8,52	1,92	6,52	2,32	4,52	2,22

The grade 8 estimates indicate a different tendency to the grade 5 and 9 estimates, in that students' self-efficacy judgments are less accurate and more biased the harder the tasks get, mirroring a linear relationship. For grades 5 and 9 students' self-efficacy judgments are the least accurate and the most biased for medium level tasks, followed by hard and then easy tasks (most accurate and least biased). These different tendencies may be related to the level of difficulty of the test relative to the grade 8 students. Students in grade 8 and grade 9 sit the same test, thus this test is more difficult for grade 8 than grade 9 students (as evidenced in my sample in the higher mean national test score for grade 9 than grade 8 students, see Table 3 in section 2.3.3). It is possible grade 8 students in general underestimated the level of difficulty of the test, and as such tended to be relatively more overconfident the harder the items were. This would indicate the different tendencies are a product of the overall test difficulty, rather than reflecting the relationship between levels of difficulty (in relative terms) and students' judgments of self-efficacy for easy, medium, and hard tasks.

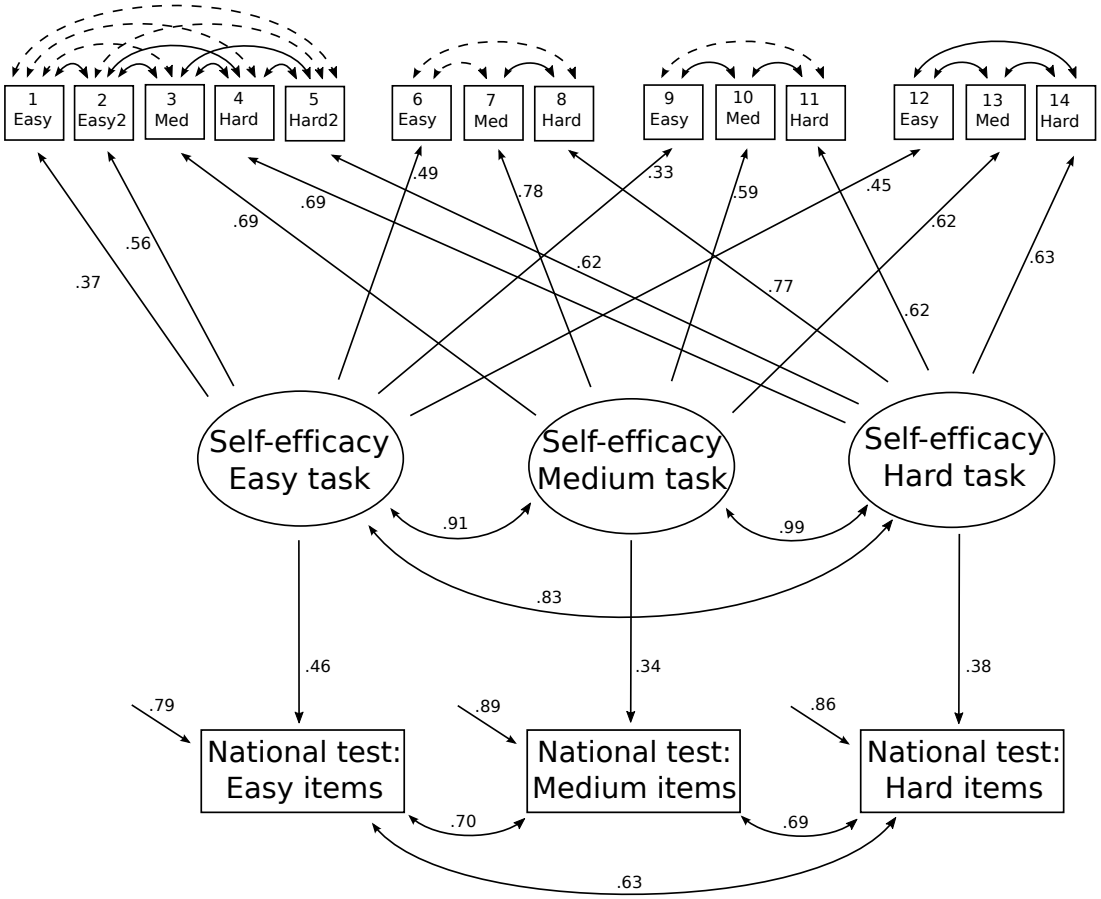
Interestingly, grade 8 students on average did better than grade 9 students on medium level test items. It is important to take a note here of the methodology used to distinguish between items of easy, medium, and hard levels of difficulty (described in section 2.3.5), and that the number of test items assigned to each category differed between the grade levels (see Tables 4-7). It is thus hard to say for sure whether the higher mean scores reflect methodological issues or true substantive differences.

While the descriptives presented above provide interesting insights and display certain tendencies, I cannot from simple inspection of the estimates be certain the differences and tendencies are statistically significant, reflecting likely "real" differences in the population. Thus, in order to investigate statistically the association between students' self-efficacy and national test items of varying difficulty, I specified structural equation models where students' scores on easy, medium, and hard items from the national tests were regressed on their self-efficacy for perceived easy, medium, and hard level tasks. These models build on the best-fitting model presented in section 2.4.3, thus I did not specify further alternative models for comparison. I used the same cut-offs as presented in section 2.3.5 to group the test items into easy, medium, and hard level tasks. Unlike my best-fitting model for research question 2 (section 2.4.3), in these structural equation models I was able to specify the relationship between national test scores and self-efficacy as regressions rather than correlations, as the self-efficacy beliefs for easy, medium, and hard level tasks were related to separate dependent variables. The fit indices from these structural equation models are presented in Table 23.

Table 23: Results from path models with matched self-efficacy and test items

No:	Model description:	χ^2	*p*	RMSEA	CFI	TLI	SRMR
1	Grade 5 (Figure 4) Easy, medium, and hard test items regressed on self-efficacy for easy, medium, and hard level tasks	$\chi^2_{(94)}$= 206.40 χ^2/df=2.20	*p*<.001	.057	.955	.935	.055
2b	Grade 8 (Figure 5) Easy, medium, and hard test items regressed on self-efficacy for easy, medium, and hard level tasks	$\chi^2_{(94)}$=174.51 χ^2/df=1.82	*p*<.001	.061	.960	.942	.097
3b	Grade 9 (Figure 6) Easy, medium, and hard test items regressed on self-efficacy for easy, medium, and hard level tasks	$\chi^2_{(94)}$= 189.52 χ^2/df=2.02	*p*<.001	.063	.963	.947	.063

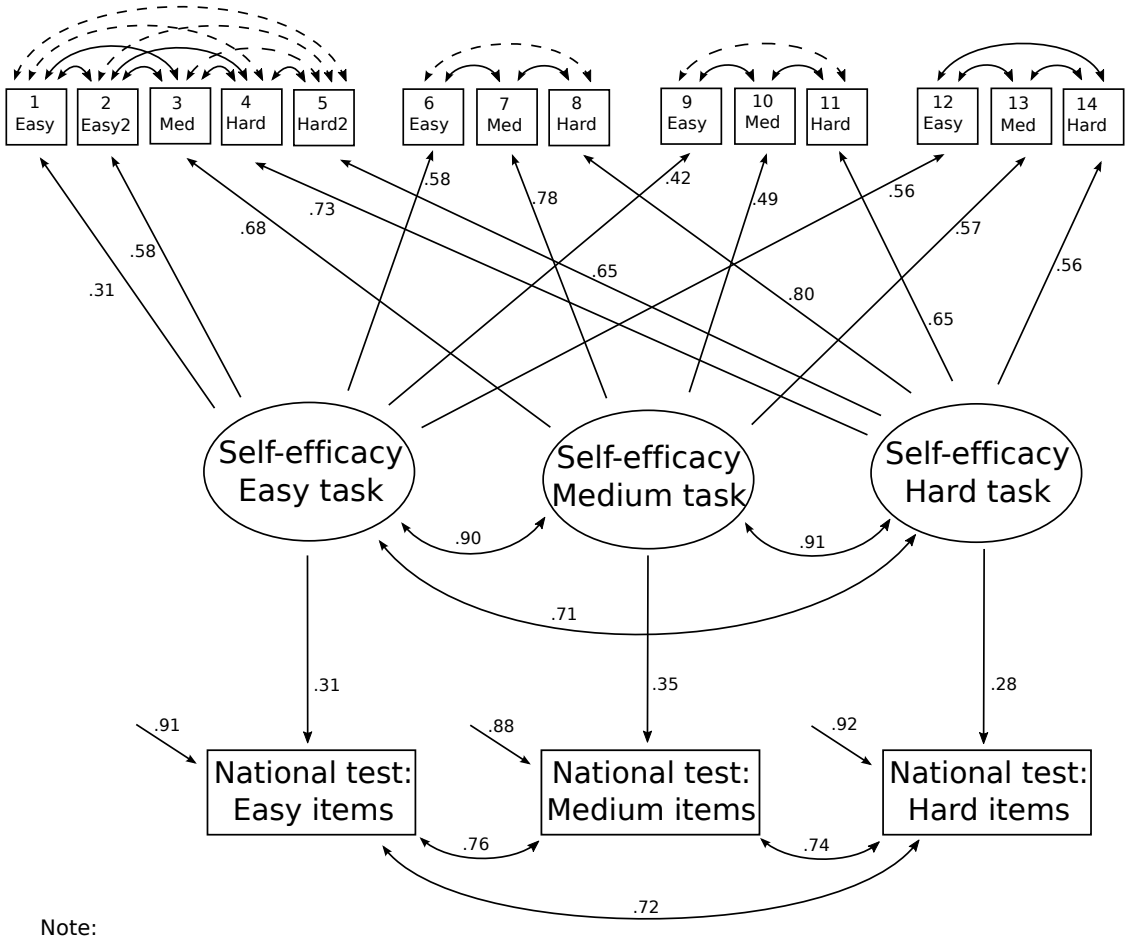
As can be seen in Table 23, the models resulted in "acceptable" fit indices for all grades. Furthermore, all three models had reasonable (low standard errors) and meaningful (in substantive terms) parameter estimates, and resulted in no negative residual variances. Parameter estimates are presented in Figures 15, 16, and 17, for grades 5, 8, and 9, respectively.



Note:
All estimates are standardized (STYX) estimates from Mplus (Muthén & Muthén 2012).
Estimates for CUs are not given due to space limitations.

Figure 15: Model with easy, medium, & hard national tests on self-efficacy, grade 5

The grade 5 path model (Figure 15) indicates that students' self-efficacy beliefs for all three levels of task difficulty predict subsequent scores on national tests in mathematics, for corresponding levels of difficulty items. The relationship is strongest between self-efficacy and test scores in relation to easy level tasks (.46), followed by hard (.38) and medium (.34) level tasks. In order to investigate the potential effect of the cohort as a covariate, I specified also a model including test items regressed on the cohort. This model indicated the cohort did not have an effect on the grade 5 national test scores. Unfortunately, a model where both national test parcels and self-efficacy constructs were regressed on the cohort variable did not converge.



Note:
All estimates are standardized (STYX) estimates from Mplus (Muthén & Muthén 2012).
Estimates for CUs are not given due to space limitations.

Figure 16: Model with easy, medium, & hard national tests on self-efficacy, grade 8

The grade 8 path model (Figure 16) also indicates that students' self-efficacy beliefs for all three levels of perceived task difficulty predict their scores on national test items of corresponding level of difficulty. However, the magnitude of the relationships differ as compared to the grade 5 results: in the case of grade 8 students, the relationship is strongest for medium level tasks (.35), followed by easy (.31) and hard (.28) level tasks. The estimates for the relationships between easy and hard levels of self-efficacy and easy and hard test scores are also lower as compared to the grade 5 estimates.

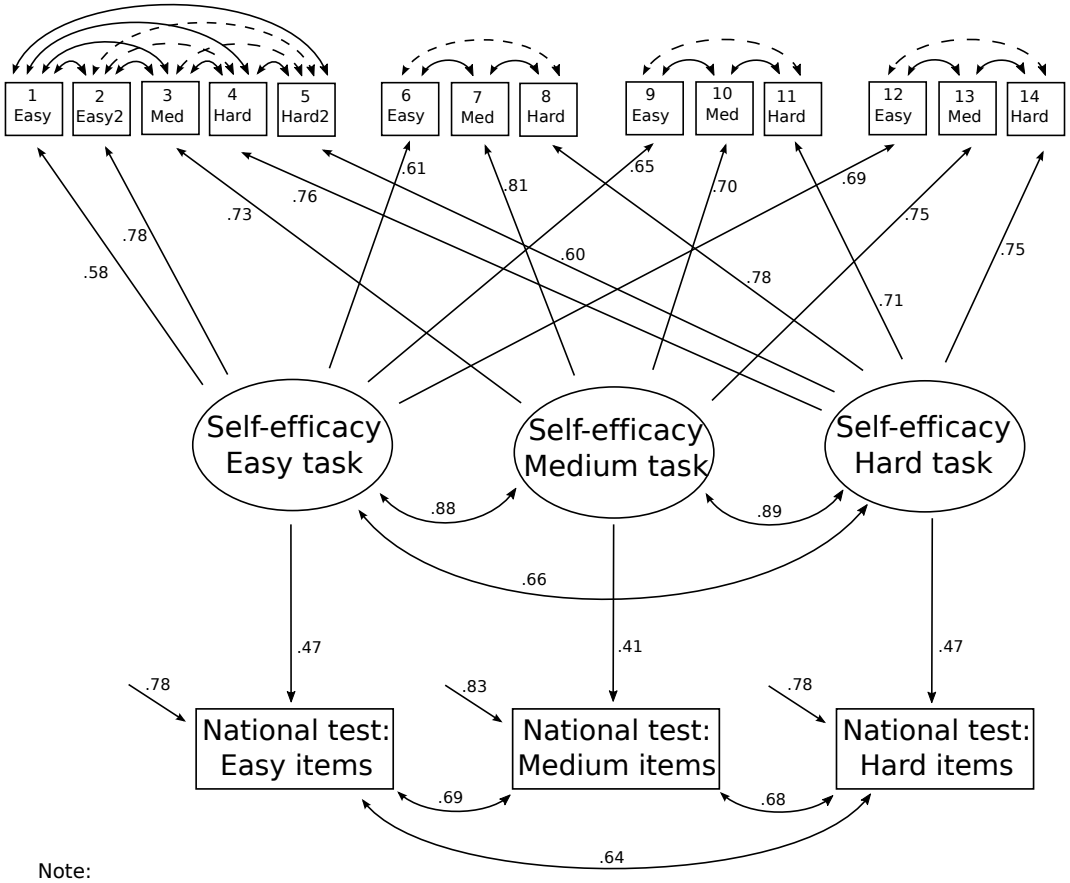


Figure 17: Model with easy, medium, & hard national tests on self-efficacy, grade 9

Figure 17 displays the path model from the grade 9 sample. Similar to the path models for grades 5 and 8, students' self-efficacy beliefs for easy, medium, and hard level tasks predict their scores on subsequent, corresponding, national test items. Similar to the grade 5 results, the relationship is stronger in relation to easy (.47) and hard (.47) level tasks than medium level tasks (.41). The magnitudes of the associations between self-efficacy and subsequent national test scores are larger in the case of grade 9 students as compared to grade 5 and grade 8 students, for all task levels.

To sum up, structural equation models for grades 5, 8, and 9 all indicate students' self-efficacy for easy, medium, and hard perceived level tasks predict their subsequent scores for easy, medium, and

hard level items on national tests in mathematics. While the grade 5 and 9 findings indicate the association between self-efficacy and test scores is weaker for medium level, as opposed to easy and hard level tasks, grade 8 results indicate the opposite.

Overall, these findings provide an interesting extension and contrast to my findings on research question 2, where I investigated the relationship between students' levels of self-efficacy and their overall national test scores. In research question 2, I found that the relationship between students' self-efficacy and their subsequent national test scores is stronger in relation to medium level of perceived task difficulty. It is likely the contrasting findings are related to the dependent variable measure, the national test scores. In research question two the *overall* national test score was used, which includes students' performance on easy, medium, and hard tasks combined. On average, students are more likely to make mistakes on hard tasks, and get the correct answer on easy tasks, while they may do a bit of both on medium level tasks. As such, we might expect students' self-efficacy for medium level tasks to best predict their overall scores on subsequent national tests, as was found in relation to research question 2. This finding might thus reflect that self-efficacy for perceived medium level difficulty tasks is more closely aligned with the performance measure (a methodological aspect), rather than that students are better at judging their performance capability for medium level difficulty tasks (a substantive aspect).

In contrast to the models used to investigate research question 2, in research question 3 I partitioned the relationships between easy, medium, and hard level tasks, estimating the effect of self-efficacy on national test scores for each of the levels of task difficulty separately. Assuming the findings from research question 3 reflect substantive rather than methodological differences, the question arises why there would be grade differences in regard to the relative strength of association between

self-efficacy and national test scores for different levels of task difficulty. Possible reasons for this will be discussed in section 2.5.3.

2.5 Discussion

In this chapter, I investigated the structural validity of the Self-Efficacy Gradations of Difficulty Questionnaire, which disentangles test taking facets and levels of difficulty from strength of self-efficacy (research question 1). Furthermore, I examined how the aforementioned constructs were associated with performance on national tests in mathematics overall (research question 2) and for easy, medium, and hard items of the test (research question 3). Next I discuss my findings, organised by research question.

2.5.1 RQ1: Structural validity of the proposed measure

According to the most common conceptualisation of self-efficacy (Bandura, 1997; Schunk & Pajares, 2010; Zimmerman, 2000), self-efficacy varies according to specificity (facets), level (levels of difficulty), and strength (confidence). The present findings indicate that it is reasonable to measure mathematics self-efficacy with a multidimensional instrument, including latent constructs representing test taking facets and levels of difficulty. Consistent with this idea, models fitted the data better when both facets and levels of difficulty were included in the model specification, through correlated uniqueness terms (Models 2b, 3b, 4b and 5b) or latent method factors (Model 6), as according to MTMM logic (Marsh et al. 2010). Models that included only facets or levels of self-efficacy resulted in poor fit indices (Models 2, 3, 4, and 5). Overall self-efficacy strength was estimated combining the scores from all items, in Model 1. The results indicate a measure that takes into account the multidimensional nature of self-efficacy is superior to a unidimensional measure of self-efficacy strength, which does not account for test taking facets or levels of perceived difficulty.

In the present study, the best-fitting measurement model is a three correlated levels model with CUs specified according to four facets (Model 3b). This indicates that the students in my study differentiated between levels of perceived difficulty of test taking facets when forming their self-efficacy. The fact that this finding was replicated for two separate student samples strengthens the generalizability of the result. Previous research has emphasised the importance of task-specificity of mathematics self-efficacy (Bong, 2001; Pajares, 1996b; Pajares & Miller, 1995). The importance of including levels of difficulty in self-efficacy assessment has been asserted (Bandura, 2006; Pajares, 1996b; Zimmerman, 2000) and demonstrated empirically (Berry et al., 1989; Locke et al., 1984), although it does not seem to have previously been tested in mathematics. The current study presents empirical data highlighting the relevance of considering levels of difficulty when measuring mathematics self-efficacy.

The best fitting model has better fit indices in the case of the older (grades 8/9) than the younger (grade 5) students, and differences in model fit between best and second best models were larger with the grade 8/9 sample. Also, the high correlation between medium and hard level tasks indicated grade 5 students did not discriminate much between these. This might all reflect that older students take more account of different levels of difficulty than younger students, supporting previous research indicating that students increasingly differentiate between motivational constructs with age (Bong 2001). Further studies are needed into the relationship between student age and self-efficacy for different levels of perceived difficulty.

In the current study, each item was related to both a facet and a level of difficulty, and for each item, strength of self-efficacy was measured on an 11-point scale. To the best of my knowledge, self-efficacy in mathematics has not previously been tested in this way. Further replications are needed to confirm the validity of this approach.

Implications

If students attend to the perceived difficulty level of the task to the same extent they attend to the type of mathematical task, this indicates teachers should ensure students have experiences working on different level difficulty tasks. The task difficulty should be considered relative to students' skill levels, and the students would benefit from teacher support to interpret their experiences from engaging with tasks of different levels of difficulty. For example, students may after experiences working on difficult tasks "decide" they can only do easy tasks, which could lead them to decrease their efforts and set lower goals for more difficult tasks, which again could lead to lower performance, falsely confirming their low expectations. In this case, it could be important for the teacher to assist students in interpreting their experiences on the difficult task in the first place, and to support the student in regard to the kinds of strategies he /she might employ when working on difficult tasks.

2.5.2 RQ2: Multidimensional self-efficacy and national test scores in mathematics

I specified structural equation models on the basis of the measurement models tested for in research question 1. I proposed a theoretical model (see Figure 2), which was operationalised in a second order latent constructs measurement model including latent facets and latent levels (Model 7, see Figure 13). This model did not converge as measurement model for the present samples and resulted in non-meaningful parameter estimates when specified as structural equation model. It is rare to encounter models with more than one negative residual variance, and recommendations in this regard are inconclusive. While the model did not converge to a proper solution for the present samples this is not evidence the model is "wrong". This model should be tested in future studies, as it could provide information about the relationship with performance scores for facets and levels of self-efficacy.

In the current samples, a MTMM model in which students' scores on a national test in mathematics are related to three latent factors representing self-efficacy for different levels of difficulty fit data best (Model 3b). CUs were specified according to four test taking facets. The present findings indicate there is a relationship between level of students' self-efficacy and their scores on national tests in mathematics. In both grade 5 and grades 8/9, stronger self-efficacy is related to higher national test scores. This is in support of previous research (e.g. Multon et al., 1991; Pajares, 1996b; Stajkovic & Sommer, 2000). It appears that the relationship between self-efficacy and national test scores is moderated by age, where older students' self-efficacy is more strongly related to their national test scores than younger students' self-efficacy. This is also in line with previous research findings (Bong 2001).

In my sample, the relationship between self-efficacy and overall national test scores is strongest for self-efficacy related to medium levels of perceived difficulty. This finding supports the findings from Locke et al. (1984), as opposed to those from Chen and Zimmerman (2007). The current design mirrors Locke et al. (1984), in that levels of self-efficacy are included in the measure a priori. Furthermore, a single overall performance measure was used in the study by Locke et al. (1984) as well as to answer research question 2 in my study, unlike Chen and Zimmerman (2007), who calculated students' performance scores for different level difficulty tasks. This highlights the importance of considering the approaches to measurement when comparing substantive findings.

Findings from the younger and older samples in my study differ in regard to the relationship between self-efficacy for easy and hard tasks, and overall national test scores. In grade 5 hard tasks were the second strongest correlation of national test scores, while for grades 8/9 easy tasks were the second strongest correlate. This may be related to grade 8/9 students' test experiences in grade

5, or to the increased difficulty of the test in grade 8/9. Further investigations are needed into the relationship between levels of self-efficacy and performance outcomes for different age groups.

Implications

If students' self-efficacy is more strongly related to test scores for medium (as compared to easy or hard) level tasks, this could be an indication that students are more biased (toward under/ over-confidence) and less accurate in relation to easy and hard tasks. If this is the case, teachers might want to support students' self-efficacy for easy and hard level tasks, for instance through teaching students to break difficult tasks into several more manageable steps.

Alternatively, the above finding could be related to the current methodology, of using the overall score of the national test as the dependent variable. As mentioned in section 2.4.4, a strong relationship between self-efficacy for perceived medium level tasks and the overall national test score might in this case reflect a closer alignment between self-efficacy and performance scores for perceived medium level difficulty tasks, as compared to easy and hard level tasks. In order to disentangle the relationship between self-efficacy and test performance for various levels of difficulty tasks, from the relationship between level-specific self-efficacy with overall performance scores, we need to separate the test items into easy, medium, and hard categories, and investigate these relationships separately. This was done in research question three in the current study, which is discussed next.

2.5.3 RQ3: Self-efficacy constructs matched with national test parcels

In research question 3 I investigated the relationship between self-efficacy and test performance for different level tasks. I wanted to disentangle any differential relationships according to level of task difficulty from the relationship between levels of self-efficacy and mean performance scores. As such, I grouped the national test items into easy, medium, and hard level item groups, according to

the percentage of students who got each item correct, according to the approach taken by Chen and Zimmerman (2007). In section 2.4.1 I described the correlation coefficients between the three latent self-efficacy level variables and the three national test parcels. These correlation coefficients indicated differential relationships between self-efficacy and national test scores for different levels of difficulty tasks, and different grade levels. Most notably, the association between self-efficacy and test scores are for grades 5 and 9 highest in relation to easy level tasks, followed by hard and then medium level tasks. For grade 8, the highest association is in relation to medium level tasks.

I proceeded to calculate students' accuracy and bias for tasks of different levels of difficulty, as a function of their performances relative to their self-efficacy. I calculated these variables in order to enable comparison with the previous study by Chen and Zimmerman (2007). In my study, bias scores indicate most students in my sample were overconfident, in relation to all task levels. This is aligned with previous research, such as the findings by Chen and Zimmerman (2007). Furthermore, accuracy descriptives indicated most of the students in my sample were more accurate at judging their capabilities for easy tasks (across all grade levels), followed by either hard tasks (in the case of grades 5 and 9) or medium tasks (in the case of grade 8).

Building on the findings from my research questions one and two I used the best-fitting measurement model, and included as dependent variables the three groups of performance scores. Overall, three structural equation models indicate students' self-efficacy for different levels of perceived task difficulty predict their subsequent performance on corresponding task items. The magnitude of the relationship differs for the different grade levels, where students in grade 9 were most accurate at judging their capabilities, followed by grade 5 and grade 8 students. Previous research has indicated the accuracy of students' self-efficacy judgment increases with age (see section 2.2.2), supported in my findings by the higher degree of accuracy in grade 9 as compared to

grade 5. This is likely related to both processes of maturation as well as increased experiences with normative feedback in school. Different to previous research, the accuracy of students' self-efficacy seemed to be lower, rather than higher, in grade 8 as compared to grade 5.

Further differences between the grades were found in relation to the relative accuracy of self-efficacy for easy, medium, and hard level tasks. In my sample, students in grade 5 and grade 9 were more accurate at predicting their performances on easy and hard, as compared with medium level tasks, demonstrated through the path estimates in Figures 4 and 6. In the case of grade 8 students, their self-efficacy was more accurate in relation to medium level, as compared with easy or hard level tasks. These contrasting findings might be related to the relative difficulty of the national test in the case of grade 8 students, as compared to the grade 5 and grade 9 students (also discussed in section 2.4.4). The differences in relative difficulty of the test is evidenced in Tables 4 to 7 (section 2.3.3), which indicate the grade 8 students found the national tests harder (difficulty value = .49 in my sample and in the national sample), than students in grade 5 (difficulty value = .61 in my samples, .59 and .56 in the national samples) or grade 9 (difficulty value = .60 in my sample, .57 in the national sample).

Chen and Zimmerman (2007) found that American and Taiwanese grade 7 students' accuracy (prediction of test score results) declined linearly with item difficulty. While my results are not consistent across all grade levels, they differ from those of Chen and Zimmerman, in that the findings do not indicate linear but indicate curvilinear trends. In a similar vein to Chen and Zimmerman, Schraw et al. (1993) reported that students were more accurate at assessing their performance on easy than hard tests (through 'postdictions', i.e. students' assessments after sitting the test). Schraw et al. (1993) argued students did not differentiate between their performance capabilities for different level difficulty tasks to the same degree they differed in actual performance

on the tasks. Given that the students performed more consistently on an easy versus a hard test, the relationship between students' assessments and actual performance was stronger for the easy test. In my study, this would mean that grade 8 students overestimated their performance capability (for a relatively harder test, as compared with grade 5 and grade 9 students), and that they overestimated their capability especially in relation to easy and hard test items.

My findings may reflect a relationship between self-efficacy accuracy and *age / cognitive maturity* (a tendency that accuracy of self-efficacy increases with age, e.g. Nicholls, 1978; Schunk and Pajares, 2010) as well as a relationship with self-efficacy accuracy and *test difficulty* (if accuracy of self-efficacy decreases with higher test difficulty, e.g. Schraw, 1993). It may be that the theoretically expected positive effect on self-efficacy accuracy associated with the increase in age from grade 5 to grade 8 is counteracted by a negative effect on accuracy of self-efficacy from the relative higher test difficulty in grade 8, as compared to both grade 5 and grade 9.

When looking at the estimates between self-efficacy and test performance for different level tasks, for the different grades (Figures 15, 16, and 17), we see that students' accuracy of self-efficacy is relatively stable across grades for medium level tasks (differences between path estimates for the different grades are relatively small: grade 5 = .34, grade 8 = .35, grade 9 = .41), as compared to easy (grade 5 = .46, grade 8 = .31, grade 9 = .47), and hard level tasks (grade 5 = .38, grade 8 = .28, grade 9 = . 47). If we accept the above interpretation that the differences in self-efficacy accuracy between the grades and levels of task difficulty are related to effects of age as well as effects of overall test difficulty, my findings indicate that students' accuracy of self-efficacy for easy and hard tasks is negatively affected by an increase in (relative) test difficulty. In contrast, students' accuracy of self-efficacy for medium level tasks appears largely unaffected. The question remains why there would be differential effects on accuracy of self-efficacy for different level difficulty tasks. Given

the lack of research on level-specific mathematics self-efficacy it is of interest to see whether the present findings would be replicated in a study including students' self-efficacy for items of varying levels of difficulty included in easy and hard overall tests in mathematics for students in the same grade level.

Caution must be taken in regard to comparing the findings from my study and those of previous studies, particularly in regard to the methodologies applied. Schraw et al. (1993) did not assess self-efficacy, but 'postdictions' after taking a test. These differ conceptually and methodologically to self-efficacy as measured in my study. Chen and Zimmerman (2007) presented students with prospective test items to rate their self-efficacy, and measured their performance on similar items. Different to Chen and Zimmerman (2007), I measured students' self-efficacy for a range of test-taking facets, and used national test scores as a measure of their performance. I would expect stronger associations between the measures if the self-efficacy items and the test items in my study had been more closely aligned.

In accordance with previous findings about gender differences in self-efficacy (see section 2.2.2), the girls in my sample had weaker self-efficacy than the boys. This was true for all grades and levels, except in association with easy level tasks in grade 9. Furthermore, girls' self-efficacy scores were consistently lower than those of the boys, despite the girls' higher absolute test scores in grades 8 and 9. Gender differences in mathematics and mathematics self-efficacy have previously been assigned to gender stereotypical beliefs (e.g. Ganley & Vasilyeva, 2014; Pajares & Valiante, 2001). I my study, the absolute scores in Tables 16-18 indicate the girls also had more *accurate* self-efficacy as compared to the boys. In regard to a related construct, Marsh, Byrne, and Shavelson (1988) found that boys had higher mathematics self-concept as compared to girls, but actually poorer grades, indicating a stronger tendency for overestimation of their capabilities among boys

than girls. In a similar vein, Pajares (1996a) found that although gifted girls achieved higher test scores than their male peers, their self-efficacy did not differ to those of the boys. Considering the overall tendency of the students in my sample to overestimate their own capabilities, the girls' lower self-efficacy (as compared to the boys' higher levels of overconfidence) proved to be more accurate. It seems that while previously perceived differences in performance between girls and boys have been shown to be inaccurate (e.g. Gibbs, 2010; Lindberg, Shibley Hyde, Peterson, & Linn, 2010), girls' own perceptions of their capability in mathematics is lagging behind.

Implications

If students are generally overconfident, but have more accurate self-efficacy in relation to easy tasks (as indicated by the correlation matrices in Tables 9 to 11 and the descriptives in Tables 16-18), as compared with medium and hard level tasks, this implies students could benefit from developing accurate self-efficacy for more difficult levels of tasks. It may be that students are more adept at judging their capabilities on easy tasks, as they have extensive practice working on easy tasks in the classroom, and can relate to previous experiences engaging with such tasks. As previously discussed, teachers would do well to assist students' interpretations of their task experiences, and to model attributions such as effort and strategy use, rather than fixed ability.

In my sample, students' accuracy of self-efficacy for medium level tasks remained relatively unchanged between grade 5, grade 8, and grade 9, as compared with the accuracy of easy and hard level tasks, which differed between the grade levels. This may be related to the relative difficulty of the mathematics test, but it may also be related to variations in my sample, or other unknown factors. The relative stability of accuracy of self-efficacy for different level tasks would have implications for researchers who want to investigate the relationship between self-efficacy and performance outcomes. It is for instance possible this relationship is stronger in relation to easy

level tasks, but more stable in relation to medium level tasks, in which case researchers would need to be careful to include relevant measures of self-efficacy and performance outcomes according to the aims of their research.

The implications for classroom practice include that in addition to supporting students' experiences with and interpretations of working on different level tasks, teachers might want to consider whether students' self-efficacy is too stable, or not stable enough, in relation to both levels and types of tasks. For instance, if students' self-efficacy fluctuates to a large degree in relation to contextual factors, they might benefit from learning about different kinds of strategies they can employ to influence their own learning and performance. On the other hand, if students' self-efficacy is overly stable in the face of changing circumstances, the teacher might want to consider the accuracy of the students' self-efficacy, and whether they are adaptive. Some students develop stable, weak, self-efficacy, which can lead to reduced effort and persistence, and thus tendencies to receive false confirmation bias. In this case, teachers might need to present students with compelling "evidence" to support their mastery experiences, as mere encouragement is unlikely to have much of an effect on students' expectations.

The accuracy and bias descriptives indicate girls in my study had less optimistic and more accurate self-efficacy than boys. If girls are more accurate at assessing their performance capability than boys, this gives rise to several important questions. Bandura (1997) held that it is beneficial to have slightly optimistic self-efficacy, but we know little about what the optimal level of optimism is, or whether this differs according to gender. If girls perform to a similar level in mathematics as boys but hold weaker self-efficacy, is accurate self-efficacy beneficial to them, or does their weaker self-efficacy (as compared with those of the boys) affect their performance in an adverse manner? In my study I included descriptives of girls' and boys' accuracy and bias of self-efficacy in order to allow

comparison with another study, however my study was not designed with a view to investigate gender differences specifically. A fruitful avenue for research might be to investigate gender differences in accuracy of self-efficacy in relation to tasks of different levels of difficulty, and how potential gender differences change or develop over time. Furthermore, gender-stereotypical beliefs may be an unsatisfactory explanation in this regard, thus future studies should investigate the potential reasons behind gender differences in accuracy of self-efficacy.

2.5.4 Limitations

The findings in my study are limited to the levels and facets I chose to include, and other studies with other aims may choose to include different facets, or levels of difficulty. Perhaps some items were too easy, indicated by high scores on the easy level items. Future studies should consider including levels of perceived difficulty for different facets of mathematics. Furthermore, with the present samples, the theoretical model led to non-convergence and improper solutions. It would be valuable to carry out a future simulation study to investigate the likely causes of this.

Students' self-efficacy was collected in a relatively long questionnaire, with the risk of "exhausting" participants, particularly the younger ones. While students were informed there were no right or wrong answers and that their teachers would not see their responses, it is possible students responded to the questionnaire in a manner they perceived to be socially acceptable ('social desirability effect'). Furthermore, although I piloted my self-efficacy measure, it is possible the piloting was not extensive enough, evidenced in my a-posteriori decision to exclude from further analyses all items related to the "Grade" facet originally included.

I chose to investigate responses from grade 8 and grade 9 students together for research questions 1 and 2. This decision was made because the students took the same test, to enable a larger sample to investigate the structural validity of my measure, as well as the relationship between self-efficacy

and national test scores. While grade 8 and grade 9 students had comparable self-efficacy in regard to strength, there were differences between the students' mean scores on national tests (the mean score was higher with the grade 9 group, as would be expected). Furthermore, previous research indicates that students' accuracy of judgments of own capability increases with age. With the analyses of research question 3, it is evident that there are differences between the accuracy of self-efficacy of grade 8 and grade 9 students in my sample, and furthermore, that these differences are related to levels of task difficulty. In light of the consistent findings across grades (similar best-fitting model in grade 5 as grades 8/9), I think it is unlikely the findings from research question 1 are affected, related to the structural validity of my proposed self-efficacy measure. However, I consider it likely that the combined grade 8/9 findings from research question 2 "hide" variations at the grade level, in regard to the strength of association between level-specific self-efficacy and national test scores. These grade-level variations were addressed in research question 3.

In a similar vein, students' self-efficacy and national test scores are in my study investigated at the group level. These groups are made up of individuals with differing characteristics, thus within each group there will be individual variation. While my findings provide interesting findings of central tendencies, it is important not to make inferences about individuals from group level analyses (or vice versa). One potentially fruitful avenue for research into students' self-efficacy is to use longitudinal data to study intraindividual changes over time (e.g. Schmitz, 2006).

A further limitation of the present study is that self-efficacy and performance measures were included at a single time point, thus any previous effects, e.g. from performance to self-efficacy, were not controlled for. This could lead to an overestimation of the relationship between these factors. However, Bandura (2012) has argued that performance does not reflect ability alone, but also previous self-efficacy, and that including unadjusted past performance scores could thus

overcontrol for ability. In study 2 I include measures of both self-efficacy and performance at two time points, in order to estimate unique effects.

In my study, the method of displaying test items to the students in order to rate their self-efficacy was not possible. Furthermore, I ranked the test items into easy, medium, and hard categories according to mean student performance, however these categories do not necessarily correspond with students' perceptions of easy, medium, and hard test items. In a future study, presenting students with specific prospective problems to solve, and asking them also to rate how difficult they perceive each problem to be, might provide different results.

Participant selection in this study was not random, but purposive, and the response rate varied a lot between classes. While I investigated whether response rates influenced my selection of best-fitting model, it is possible that other variations in the sample, such as socio-economic background, influenced my results. Finally, my study investigated specific motivational theories within specific settings. The findings are related to Norwegian grade 5, 8 and 9 students, and the extent to which they are applicable beyond this context must be assessed on a case-by-case basis. It is possible that a future replication of the study in a different socio-cultural context could provide new insights.

2.6 Conclusion

I tested a new instrument responding to the proposed multidimensionality of self-efficacy, including levels of perceived difficulty as well as facets of a national test. The results indicated that students differentiate between levels of perceived difficulty, as well as test taking facets, when forming their self-efficacy. Further, I tested the associations between self-efficacy as measured by my instrument, and national test scores. I found that levels of self-efficacy were meaningfully related to national test scores, and that the relationship with overall test scores was highest for self-efficacy in relation to tasks of medium level of perceived difficulty. The findings from research question 2 were

extended and contrasted by my final research question, where I tested the relationship between self-efficacy for easy, medium, and hard levels of perceived task difficulty, matched with performance on easy, medium, and hard test items. Self-efficacy for all levels of difficulty predicted scores on subsequent, matched, test items. The association between self-efficacy and national test scores differed according to grade level as well as task difficulty, with some indications that the association is *stronger* in relation to easy level tasks, and *more stable* in relation to medium level tasks.

Future studies should investigate further students' self-efficacy for different levels of task difficulty in relation to performance outcomes across different facets of mathematics, and whether students' accuracy of self-efficacy differs as a function of gender. Also, there is a need to investigate the processes through which self-efficacy beliefs develop and change, in order to better understand how to support the development of students' self-efficacy in the classroom.

Findings from the present study imply that researchers should consider including levels of perceived difficulty when measuring mathematics self-efficacy. Students' self-efficacy may vary not only as a function of specific tasks or facets, but according to the perceived difficulty of these tasks. Furthermore, the relationship between students' self-efficacy and their task performance might vary as a function of level of task difficulty. Teachers of mathematics may wish to consider levels of task difficulty as experienced by students, to help students develop accurate self-efficacy for future tasks of different levels of difficulty.

3 Reciprocal relationships between self-efficacy and national test performance (study 2)

Several studies have looked at either the effects of students' self-efficacy on their mathematics performance results, or the effect of previous mathematics performances on their self-efficacy. Few studies have tested the theoretically proposed reciprocal relationship between mathematics self-efficacy and performance in mathematics for school age children. Furthermore, previous studies have not included levels of difficulty, although this is an integral part of the definition of self-efficacy. This study applied a new measure of self-efficacy, which included both test taking facets (related to problem solving or students' self-regulation skills during the test) and levels of perceived difficulty, to investigate their relationships with students' performance on a national test in mathematics. Two models fitted data well, and provided partial support for the reciprocal effects model. Test performance predicted scores on subsequent self-efficacy constructs, and self-efficacy constructs were associated with scores on immediately upcoming national tests, but did not predict test performance one year later. The models provided distinct estimates of the relationships between students' test performance and their level and facet-specific self-efficacy, respectively.

The findings from study 2 serve to highlight the specificity of self-efficacy, where they were associated with an immediately upcoming test, but not for the same test a year later. In contrast, national test scores predicted gains in students' self-efficacy after a year had passed, indicating the testing situation provided a memorable event for students. If national tests have long-term effects on students' self-efficacy, this implies teachers should consider how students make sense of their mathematics test experiences, to support the development of adaptive self-efficacy.

3.1 Introduction

Self-efficacy is important because extensive research indicates it is related to student learning and performance (Zimmerman, 2000). Studies have demonstrated the effect of mathematics self-efficacy on performance results (e.g. Pajares, 1996a) and vice versa (e.g. Pampaka, Kleanthous, Hutcheson, & Wake, 2011), but the proposed reciprocal relationship between self-efficacy and performance (Bandura, 1997) has received little attention. Investigating this relationship empirically is important in order to understand the relationship between self-efficacy and performance in mathematics, and the process by which self-efficacy may influence performance in mathematics and vice versa. Few studies have investigated reciprocal effects between self-efficacy and mathematics performance with longitudinal data, and none that I know of in relation to national tests in Norway.

Bandura (1997) argues self-efficacy varies according to three dimensions (specificity, strength, and level), but I am aware of no previous studies that have included self-efficacy *level of difficulty* to investigate reciprocal effects with mathematics. As I argued in study 1 (presented in chapter 2), including level of self-efficacy is important as it is an integral part of the definition of self-efficacy (Bandura, 1997). I aimed to address this research gap by testing a reciprocal effects model including students' level and facet-specific self-efficacy, and their scores on a Norwegian national test. To this aim, I applied a recently developed measure of mathematics self-efficacy that includes four facets of test taking as well as three levels of perceived difficulty. I included students' self-efficacy responses and test scores as they progressed from grade 8 to grade 9.

3.2 Theoretical background

Self-efficacy beliefs are judgments about individuals' perceived capability to perform on future tasks (Bandura, 1997), for instance students' expectations they are able to carry out the various facets involved with taking a mathematics test. Examples are problem-solving skills such as solving a certain number of problems or solving tasks of a certain challenge, or skills in self-regulation such as concentrating for a length of time or persevering through difficult problems.

Self-efficacy affects future performance through mediating processes, influencing students' behaviours and motivations. Self-efficacy may influence individuals' tendencies to approach learning tasks (Pampaka et al., 2011), their goal setting and self-regulative processes while engaged in such tasks (Zimmerman, Bandura, & Martinez-Pons, 1992), their persistence, and performance outcomes (Pajares & Miller, 1995). Self-efficacy may further influence individuals' tendencies to approach learning tasks, their effort and persistence while engaged in such tasks, as well as their self-regulatory processes (Zimmerman, 2000). In their analyses of 39 studies Multon, Brown and Lent (1991) found overall positive and statistically significant relationships between self-efficacy

and academic performance, with a mean effect size of .38. Effect sizes varied according to study characteristics, where for instance the positive relationship between self-efficacy and achievement outcome was weaker for elementary school students than high school or college students. Furthermore, Pajares and Miller (1995) argued the relationship between self-efficacy and performance varies according to the specificity of the measure (more specific = higher correlation), and whether the measure of self-efficacy is well matched (aligned) with the performance task.

The relationship between mathematics self-efficacy and performance outcomes has been investigated in a range of countries. Schulz (2005) investigated results from the Programme for International Student Assessment (PISA) 2003, which included 15 year-olds in 30 OECD (The Organisation for Economic Co-operation and Development) countries. Self-efficacy was consistently positively correlated with mathematics literacy (OECD median= .52, Norway= .55), and negatively correlated with mathematics anxiety (OECD median= -.34, Norway= -.50). Furthermore, self-efficacy has been found to predict mathematical problem-solving after controlling for factors such as self-concept, perceived usefulness of mathematics, prior experience with mathematics, cognitive ability, mathematics Grade Point Average, anxiety, and gender (Pajares, 1996a; Pajares & Miller, 1994).

There is also empirical support for the influence of past performances on self-efficacy (e.g. Pampaka et al., 2011). According to Bandura (1997), self-efficacy is formed through four sources, where mastery experiences provide the strongest source. Mastery experiences stem from individuals' appraisals of previous performance situations, for instance their experiences from a previous, similar test. Importantly, previous experiences do not influence self-efficacy directly, but are interpreted and made sense of by the individual (Usher, 2009). Thus, a performance experience will not have the same effect on the self-efficacy of two different students, as they are likely to

differ in their cognitive appraisal of this experience (for instance, was the test easy or difficult, did I receive much help, how much effort did I put in) (e.g. Yetkin Özdemir & Pape, 2013). Pampaka and colleagues (Pampaka et al., 2011) measured self-efficacy as a learning outcome of post-compulsory mathematics students. They found mathematics self-efficacy was positively related to both prior attainment as measured by students' General Certificate of Secondary Education (GCSE) mathematics grade (Pearson`s *r* ranging from .14 to .34) and subsequent achievement outcome as measured by their advanced-subsidiary mathematics grade (Pearson`s *r* ranging from .28 to .39). Further, Matsui, Matsui and Ohnishi (1990) used students' self-reported final year high school mathematics grades as proxy for mastery experiences. The authors reported these previous performances made an independent contribution to their model (R^2=.05, F(1,157) = 12.70, p < .01), and were the strongest of the four types of sources predicting self-efficacy.

A mutually reinforcing pattern of effects between self-beliefs and academic achievement is supported for several types of self-beliefs (Valentine & DuBois, 2005). Marsh (2013) described the competing self-enhancement (self-beliefs are a cause of subsequent achievement) and skill development (self-beliefs are effects of previous achievement) models, and argued instead for a reciprocal effects model in the case of self-concept. The reciprocal effects model is theoretically proposed in relation to self-efficacy (Bandura, 1997), however only few studies have investigated reciprocal effects between self-efficacy and mathematics performance empirically. Williams and Williams (2010) used cross-sectional data to investigate reciprocal effects using PISA 2003 results, arguing that it is not necessary to include causal lag (time interval) in structural equation models, when assumptions of system equilibrium are met. That is, their model was based on the assumption that the 15 year-old participants had reached realistic self-appraisals, through feedback from previous experiences. Williams and Williams (2010) reported that their model fit the data well for PISA data from 26 out of 33 countries. They argued a reciprocal effects model was supported, and

that this might be a fundamental psychological process transcending cultural and national borders. At the same time reciprocal effects might not work the same way in all countries (Williams & Williams, 2010). In the study by Williams and Williams (2010) the data from Norway was excluded from the analyses due to little variation in the grade-level variable. Hannula et al. (2014) investigated the dominant direction of effect between mathematics performance and self-efficacy, measured every three years as students progressed from grade 3 to grade 9 in Finland. Their results supported a reciprocal effects model, and indicated that the dominant direction of the effect was from performance to self-efficacy (grade 3 performance to grade 6 self-efficacy β = .34, grade 6 performance to grade 9 self-efficacy β = .30). However, the estimates for the paths from self-efficacy to performance differed markedly (grade 3 self-efficacy to grade 6 performance β = .16, grade 6 self-efficacy to grade 9 performance β = .26), and the second estimate for the effect of self-efficacy on performance was comparable to that of the second estimate of the effect of performance on self-efficacy. Both measures were (also) relatively stable over time (self-efficacy on self-efficacy β = .36, β = .49, performance on performance β = .64, β = .65). The authors noted the effect from self-efficacy on performance was larger for the older students in the sample.

Measures of self-efficacy have typically included self-efficacy strength and self-efficacy specificity (also called generality). While Bandura proposes a third dimension of self-efficacy (1997), only few studies have included level of difficulty. Students' self-efficacy may vary between different facets of mathematics, and in regard to different perceived levels of difficulty. Pajares and Graham (1999) found that grade 6 students' self-efficacy for solving problems predicted their test performance (the test involved solving similar problems) both the following day and one year later, while self-efficacy for self-regulated learning did not. Locke et al. (1984) investigated self-efficacy in relation to the task naming uses for common objects, and found that self-efficacy for tasks of medium difficulty were most predictive of subsequent task performance (r with performance ranging from

.40 to .53 over three trials), followed by self-efficacy for easy tasks (*r* ranging from .30 to .39) and very hard tasks (*r* ranging from .19 to .40). The authors also found support for reciprocal effects between self-efficacy and performance. They conducted a path analysis, and found self-efficacy was significantly predicted by past performance (β = .31), as well as predicted future performance (β =20).

3.2.1 Theoretical model and research questions

In line with theory and empirical findings discussed, I propose a theoretical model (see Figure 18), where self-efficacy and mathematics performance are reciprocally related. In the current study, I included Norwegian students' self-efficacy responses and subsequent test scores from the beginning of grades 8 and 9 to investigate this model. In the model, students' self-efficacy in grade 8 and grade 9 is related to their performance in mathematics in the same year (this relationship is directional, from self-efficacy to national test scores, given that students fill out self-report questionnaires prior to taking national tests). Furthermore, students' mathematics performances in grade 8 predict their self-efficacy in grade 9. Both self-efficacy and mathematics performance are relatively stable (mathematics performance the most stable of the two), with the grade 8 scores predicting scores in grade 9. I further propose different self-efficacy constructs (level and facet-specific self-efficacy) have differential relationships with mathematics performance.

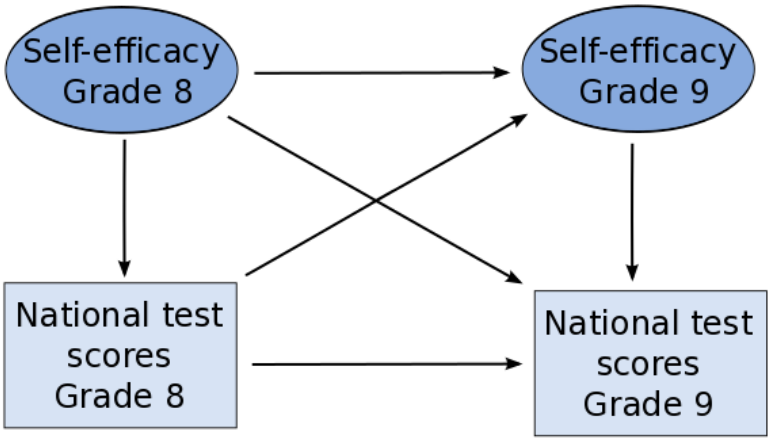


Figure 18: Theoretical model for study 2

The research questions I address in this chapter are:

1) *What is the relationship between self-efficacy and performance on national tests in mathematics in grade 8, and self-efficacy and performance on national tests in mathematics in grade 9?*
2) *Do the relationships between self-efficacy and performance on national tests differ according to levels or facets of self-efficacy?*

Hypotheses, research question 1:

1) Overall, there is a reciprocal relationship between self-efficacy and performance results, where self-efficacy predicts performance results and performance results predict self-efficacy.
2) Performance results in grade 8 will predict change in self-efficacy from grade 8 to grade 9 (the "skill development model", e.g. Skaalvik & Valås, 1999).
3) Self-efficacy in grade 8 will predict change in performance from grade 8 to grade 9 (the "self-enhancement model", e.g. Bong & Skaalvik, 2003).
4) Self-efficacy and performance results are associated in both grade 8 and grade 9.

5) The relationship between self-efficacy and (immediate) performance results is stronger than the relationship between self-efficacy and performance results one year later (taking temporal dissociation into account).
6) There is relative stability across time for both self-efficacy and performance results, where performance results are more stable than self-efficacy.

Hypotheses, research question 2:

7) The relationship between self-efficacy and performance results differs between levels of self-efficacy, where medium and hard levels of self-efficacy are more strongly associated with performance results than the easy level (e.g. Locke et al., 1984).
8) The relationship between self-efficacy and performance results differs between facets of self-efficacy, where facets related to problem solving self-efficacy are more strongly associated with test scores than are facets related to self-efficacy for self-regulation. (e.g. Pajares & Graham, 1999).

3.3 Methodology

3.3.1 Participants

The participants were 95 students (44 female) in Norwegian secondary school that completed self-report questionnaires and took national tests in mathematics at the beginning of grades 8 and 9 (13 and 14 years old). The participants were part of a larger sample, selected for cross-sectional investigations (see section 2.3.1, or Sørlie & Söderlund, 2015). Included in the study were schools where students had performed above and below what might be expected on national tests, considering measures of socio-economic-status. Each year, students responded to a questionnaire about their self-efficacy for an upcoming national test in mathematics. The questionnaire was distributed and collected by the mathematics teachers, who received detailed instructions regarding

the procedure, including to answer questions for clarification purposes, but to ensure students fill out the questionnaires independently. For a discussion of ethical considerations please see section 2.3.2 of study 1.

3.3.2 Measures

The Self-efficacy Gradations of Difficulty Questionnaire (see Appendix I) was applied to measure students' self-efficacy at the two time points (see study 1 for a detailed analysis of this measure). This is a recently developed multidimensional measure of mathematics self-efficacy, which includes four test-taking facets (facet-specific self-efficacy) related to problem solving (complete a number of problems, solve tasks of a certain challenge) and self-regulation (concentrate, not give up), and three levels (easy, medium, and hard) of perceived difficulty (self-efficacy level) (see study 1, section 2.5.1). Each of the 14 items in the measure are related to one test taking facet and one level of difficulty within each facet. For each item, students are asked to indicate their confidence (strength of self-efficacy) on an 11-point scale from 0 "not at all certain" to 10 "highly certain". An example is the facet "Concentrate". The description reads "Some people can find it hard to concentrate when sitting tests. How certain are you that you will be able to concentrate during the national tests in mathematics?". The first item related to this facet is "During the national tests this year I will be able to concentrate: at least a little while", while the two items of increasing difficulty level end with ": at least half the test", and ": the whole test". The structural validity of this measure was tested (see study 1, presented in chapter 2), and the resulting best-fit measurement model included three latent level constructs, with correlated uniquenesses specified for each of the four facets.

The performance measure was raw scores from national tests in numeracy (Norwegian Directorate for Education, 2016). "Numeracy" is similar to what researchers generally refer to as mathematics. Norwegian students sit these tests at the start of the school year in grades 5, 8 and 9. Students in

grades 8 and 9 sit the same test, which involves 58 problems, scored as either correct (1) or incorrect (0). I used unique identifiers to link the national test scores with the questionnaire responses. Item descriptives are presented in Table 24.

Table 24: Item descriptives

Descriptives

	Item	n	missing	M	SD	Range	Skew	Kurt
1	No. of problems_easy	92	3 (3%)	10.93	.36	3	-6.83*	52.06*
2	No. of problems _easy2	92	3 (3%)	10.60	.94	5	-2.78	8.32
3	No. of problems _med	93	2 (2%)	9.22	2.09	7	-.86	**-.52**
4	No. of problems _hard	93	2 (2%)	6.71	2.55	9	**-.07**	-1.07
5	No. of problems _hard2	91	4 (4%)	3.41	2.61	9	1.03	**.08**
6	Solve tasks_easy	92	3 (3%)	10.17	1.19	5	-2.02	4.49
7	Solve tasks_med	93	2 (2%)	7.88	1.83	10	-.45	**.68**
8	Solve tasks_hard	92	3 (3%)	5.10	2.17	8	**.06**	-1.03
9	Concentrate _easy	92	3 (3%)	10.61	1.45	10	-4.69*	24.71*
10	Concentrate _med	92	3 (3%)	9.76	1.70	6	-1.24	**.38**
11	Concentrate _hard	93	2 (2%)	7.40	2.83	10	-.52	**-.62**
12	Not give up _easy	92	3 (3%)	10.05	1.65	10	-2.77	10.15*
13	Not give up _med	93	2 (2%)	8.62	1.92	7	**-.38**	**-.77**
14	Not give up _hard	92	3 (3%)	7.17	2.45	9	**.06**	-1.05
15	National test score	95	0	30.61	11.18	48.00	**-.03**	-.85
16	t2No. of problems_easy	89	6 (7%)	10.85	.61	4	-4.75*	23.69*
17	t2No. of problems _easy2	89	6 (7%)	10.39	1.40	7	-2.82	8.13
18	t2No. of problems _med	89	6 (7%)	9.11	2.24	9	-1.09	**.39**
19	t2No. of problems _hard	89	6 (7%)	6.40	2.67	10	**-.10**	-.92
20	t2No. of problems _hard2	89	6 (7%)	3.34	2.68	10	1.10	**.18**
21	t2Solve tasks_easy	89	6 (7%)	10.18	1.34	5	-2.12	4.12
22	t2Solve tasks_med	89	6 (7%)	8.01	2.08	10	-.72	1.24
23	t2Solve tasks_hard	89	6 (7%)	5.03	2.61	9	**.14**	-.91
24	t2Concentrate _easy	89	6 (7%)	10.56	1.22	7	-3.53*	13.37*
25	t2Concentrate _med	89	6 (7%)	9.57	2.06	7	-1.27	**.36**
26	t2Concentrate _hard	89	6 (7%)	7.25	3.10	10	-.45	-1.14
27	t2Not give up _easy	89	6 (7%)	9.87	1.87	10	-2.11	5.25
28	t2Not give up _med	89	6 (7%)	8.53	2.17	10	-.65	**.13**
29	t2Not give up _hard	89	6 (7%)	6.73	3.04	10	**-.19**	-1.14
30	t2National test score	93	2 (2%)	35.69	11.03	47.00	-.58	**-.46**

Note: Skewness /kurtosis z-score is larger than 1.65 for all estimates, *except* those in bold

*Absolute value higher than recommended cut-off (Kline, 2011, p.63)

3.3.3 Specification of empirical models

My modelling choices were informed by the previously established factor structure of the measure, as well as my theoretical model. In study 1 (presented in chapter 2) I found that the best-fit models were those that accounted for the multidimensional nature of self-efficacy, through correlated latent constructs and correlated uniquenesses. In order to estimate the relationships between latent constructs in my models I used item parceling technique.

According to Marsh et al. (1998, p. 184) the logic behind using item parcels is that "each parcel is likely to be more strongly related to the latent factor, is less likely to be influenced by the idiosyncratic wording and method effects associated with individual items, <and> is more likely to meet the typical assumptions of normality in maximum likelihood approaches to CFA". Furthermore, models based on parcels result in a greater number of proper solutions, and often better fit, than the corresponding item-level solution (Marsh et al., 1998). While popular, the technique of parceling has been debated. Bandalos and Finney (2001) argued item-parceling can lead to misleading results, obscuring the underlying factor structure of the measure applied, for instance items loading on two factors. If researchers are concerned with understanding the relations among items in a measure, parceling may lead to misleading results and biased factor loadings. However, when researchers are interested in estimating the relations among latent variables, parceling may aid rather than hinder this aim (Little, Cunningham, Shahar, & Widaman, 2002). According to Little et al. (2002, p. 169) "Eliminating the residual through parceling is as effective as explicitly correlating the residual.". Regardless of the reason for parceling, the original factor structure of the measure applied should be established and reported (Bandalos and Finney, 2001; Little et al, 2002). In study 1 (see chapter 2) I provided support for both the facet-structure and the level-structure of the multidimensional measure.

Bong and Hocevar (2002) used item parcels when investigating convergent and discriminant validity of several self-efficacy measures, applying multi-trait multi-method data. The authors listed the following three reasons for using item parcels: small sample size (theirs was 358 participants in study 1 and 235 participants in study 2), that parcels are known to meet the normality assumptions of structural equation modeling, and that their measures appeared "relatively unidimensional" (p.149).

While my measure of self-efficacy is multidimensional, the factor structure has been tested and established in a previous study (see study 1, presented in chapter 2). Item parcels were formed to achieve "clean" latent constructs from each year, through aggregating items in such a way that the secondary loading was spread across parcels. For example, the "easy" parcel contains all items related to the easy latent construct, across four facets. Similarly, the "concentrate" parcel includes all items related to the concentrate facet, across three levels of difficulty. In this manner, the effects from the secondary construct should be counterbalanced (Little et al., 2002).

Two reciprocal effects models were specified. The hypothesized model (Model 1) includes national test scores and three item parcels for self-efficacy levels, representing the best-fit model from study 1 (see Figure 14, section 2.4.3). The alternative model (Model 2) includes national test scores and four item parcels for facet-specific self-efficacy. In both models, self-efficacy constructs in grade 8 are related to test scores in grade 8. Similarly, self-efficacy constructs in grade 9 are related to test scores in grade 9. Self-efficacy constructs are intercorrelated in grade 8, and in grade 9. Furthermore, corresponding self-efficacy constructs are related across the two time points (e.g. self-efficacy easy in grade 8 predicts self-efficacy easy in grade 9). Test scores in grade 8 predict test scores in grade 9. Two final relationships were tested: the paths from test scores in grade 8 on self-

efficacy constructs in grade 9, and the paths from self-efficacy constructs in grade 8 on test scores in grade 9.

3.3.4 Analyses

Similar to study 1, I analysed the data with structural equation modeling, using the robust maximum likelihood (MLR) estimator in Mplus (version 1.31 for Mac: Muthén & Muthén, 2012). In this estimator all available data is used (Graham & Coffman, 2012). For the correlation coefficients I used pairwise deletion in SPSS. I used fit indices recommended by previous studies (see section 2.3.5) (e.g. Morin et al., 2014; Schermelleh-Engel et al., 2003), specifically I used the chi square (χ^2/df =<3 acceptable), the Root Mean Square Error of Approximation (RMSEA <.08 acceptable), the Standardized Root Mean Square Residual (SRMR<.10 acceptable), the Comparative Fit Index (CFI) and the Tucker-Lewis index (TLI) (CFI/TLI > .90 acceptable). To assess improvement in fit between models, I used the following cut-offs: ΔRMSEA (.015), ΔCFI (.010), ΔSRMR (.030).

3.4 Results

3.4.1 Correlation coefficients

As can be seen in the correlation matrix (see Table 25), item cross-correlations for items from the same year are consistent with previous results from the same measure (see section 2.4.1), indicating a multidimensional structure (in case of the grade 8 items this is not surprising, given that the current study uses data from a subsample of participants from study 1). Item correlations are generally high between items associated with adjoining levels of difficulty (easy with medium, medium with hard) within the same facet. Similarly, across facets, there is a tendency that items of similar level of difficulty have higher cross-correlations than items associated with different levels. These patterns of cross-correlations indicate items are related to two types of latent constructs: levels (easy, medium, and hard) and facets (no. of problems, solve tasks, concentrate, and not give up). Furthermore, within each year, all non-significant estimates are in association with easy level

items. For instance, the "concentrate_easy" item is associated with seven non-significant correlations in grade 8 (item number 9) and four non-significant correlations in grade 9 (item number 24).

When looking at the autocorrelations between items in grade 8 and the corresponding (same) item in grade 9, two tendencies are evident. First, autocorrelations are significant for all items, except for two cases, both associated with the easy level (no. of problems_2, and concentrate_easy). Second, the magnitudes of the associations are consistently weaker for items associated with the easy level, than the medium and hard levels. To illustrate this: the strongest autocorrelation among the easy level items is .21, while the weakest autocorrelations among the medium and hard level items is .26 and .29, respectively.

While the cross-correlations for each of the years indicate support for a multidimensional construct, the autocorrelations indicate that students' scores on the medium and hard level items were more stable across time, than students' scores on the easy level items. The non-significant autocorrelations between the grade 8 and grade 9 easy level items might be related to a lack of variability in scores. Most students were highly confident in relation to the easy level items (evident through high mean scores and small standard deviations on these items, see item descriptives in Table 24, presented in section 3.3.2).

Table 25: Correlation matrix of self-efficacy grade 8 and grade 9

Zero-order correlation coefficients among measures

Measure	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
1 No. of problems_1																													
2 No. of problems _2	.64**																												
3 No. of problems _3	.32**	.61**																											
4 No. of problems _4	.17**	.41**	.76**																										
5 No. of problems _5	.13**	.24**	.46**	.72**																									
6 Solve tasks_easy	.39**	.28	.33**	.26**	.24**																								
7 Solve tasks_med	.18**	.33**	.57**	.59**	.46**	.53**			Grade 8 with grade 8 items							(Bottom left corner: Grade 8 with grade 9 items)													
8 Solve tasks_hard	.17*	.38**	.51**	.67**	.63**	.23**	.73**																						
9 Concentrate _easy	.27	.13	.14	.12	.05	.41*	.37**	.09																					
10 Concentrate _med	.26*	.29*	.31**	.33**	.29**	.50**	.50**	.30**	.59**																				
11 Concentrate _hard	.16	.39**	.35**	.41**	.39**	.39**	.50**	.50**	.17	.65**																			
12 Not give up _easy	.41**	.22	.19	.16*	.10	.32*	.29*	.08	.28*	.36**	.30**																		
13 Not give up _med	.36**	.29*	.38**	.34**	.26**	.24*	.42**	.31**	.27*	.51**	.47**	.70**																	
14 Not give up _hard	.26**	.28**	.36**	.41**	.38**	.17	.51**	.51**	.19*	.45**	.49**	.36**	.80**																
15 National test score	.07	.25**	.39**	.34**	.19	-.02	.37**	.34**	.23**	.24*	.16	.29**	.32**	.26**															
16 t2No. of problems_1	-.05**	-.01	.07	.16	.16**	.13	-.05	-.07	.16	.16	.12	.17	.00	-.16*	.21**														
17 t2No. of problems _2	.13	.16	.29*	.25*	.20**	.26	.13	.10	.18	.28*	.21*	.03	.07	.01	.29**	.64**													
18 t2No. of problems _3	.19**	.21*	.39**	.27**	.24**	.25*	.31**	.29*	.16	.23*	.22*	.16	.22*	.21*	.32**	.35**	.69**												
19 t2No. of problems _4	.20**	.20*	.42**	.42**	.32**	.21*	.40**	.36**	.16*	.27**	.31**	.15	.29**	.36**	.28**	.26**	.55**	.80**											
20 t2No. of problems _5	.14**	.06	.20	.28*	.29*	.11	.31**	.31**	.16**	.28**	.28**	.18*	.26**	.30**	.15	.13*	.31**	.56**	.80**										
21 t2Solve tasks_easy	-.05	-.11	-.04	-.05	.07	.20*	.04	-.12	-.03	.06	.03	-.06	-.09	-.15	-.09	.10	.32*	.31**	.25**	.24**			Grade 9 with grade 9 items					
22 t2Solve tasks_med	.09	.02	.23*	.25*	.28**	.17	.29**	.24**	-.04	.31	.22**	.15	.22*	.21*	.22*	.17**	.40**	.54**	.60**	.59**	.57**								
23 t2Solve tasks_hard	.16**	.07	.22*	.28**	.20*	.01	.30**	.36**	.13	.29**	.24*	.22**	.27**	.33**	.33**	.17*	.27**	.48**	.64**	.70**	.09	.66**							
24 t2Concentrate _easy	-.07**	-.01	.01	.13	.14*	.31**	.11	.06	.02	.21*	.11	-.04	.01	.03	-.03	.28	.38**	.19	.23**	.12	.50**	.28**	.06						
25 t2Concentrate _med	.12*	.21*	.16	.17	.16	.28*	.15	.11	.16	.44**	.33**	.09	.20*	.25*	.07	.33**	.40**	.27*	.29**	.25**	.36**	.39**	.29**	.70**					
26 t2Concentrate _hard	.17**	.26**	.21	.19	.19*	.21*	.15	.13	.11	.41**	.38**	.20**	.23*	.26**	.10	.33**	.40**	.41**	.42**	.40**	.28**	.45**	.47**	.39**	.82**				
27 t2Not give up _easy	.13	.14	.34**	.36**	.24**	.31**	.32**	.19**	.11	.35**	.16*	.21*	.23**	.13	.28**	.27*	.53**	.42**	.45**	.28**	.53**	.58**	.33**	.48**	.51**	.47**			
28 t2Not give up _med	.13*	.16	.29**	.22*	.17*	.16	.15	.17	.03	.27**	.23**	.14	.26**	.22*	.26**	.29**	.44**	.54**	.53**	.42**	.40**	.59**	.46**	.39**	.64**	.65**	.67**		
29 t2Not give up _hard	.16**	.17*	.28**	.22*	.15	.06	.15	.26**	.08	.20*	.23*	.11	.29**	.34**	.31**	.22	.34**	.50**	.55**	.48**	.18*	.48**	.58**	.23**	.52**	.65**	.41**	.84**	
30 t2National test score	.08	.12	.39**	.33**	.20	.06	.34**	.28**	.24*	.23*	.14	.29**	.32**	.30**	.87**	.21*	.32**	.39**	.31**	.20	-.01	.27**	.34**	.02	.16	.17	.30**	.38**	.42**

*Correlation is significant at the p<.05 level (2-tailed) **Correlation is significant at the p<.01 level (2-tailed)

3.4.2 Structural equation models

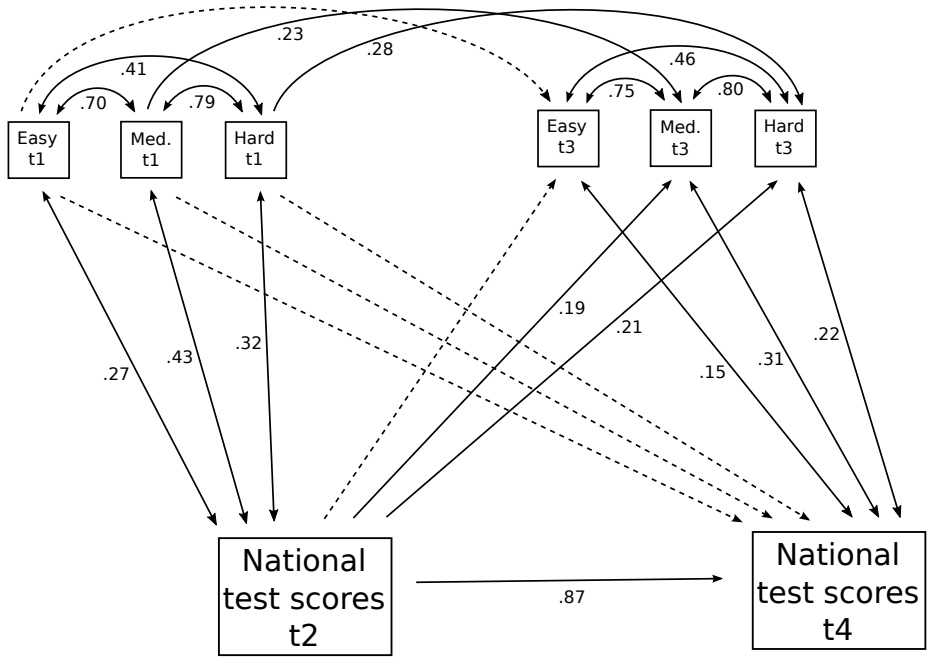
Fit indices from the structural equation models are presented in Table 26. Both the proposed models resulted in excellent fit, reasonable parameter estimates, and small standard errors. The exception is the easy level construct in Model 1, which was associated with large standard errors in both years (the standard error of the mean of the year one construct is 2.53, and of the intercept of the year two construct is 1.90, both STYX standardization). The standard errors in the Model 2 results are all smaller than .11, STYX standardization.

Table 26: Fit indices from structural equation models

	Model	χ^2	*p*	RMSEA	CFI	TLI	SRMR
1	Correlated levels model, reciprocal relationship with test scores	$\chi^2_{(6)}$= 4.397 χ^2/df=0.73	*p*=.623	.000	1.00	1.02	.048
2	Correlated facets model, reciprocal relationship with test scores	$\chi^2_{(12)}$= 10.890 χ^2/df=0.91	*p*=.538	.000	1.00	1.01	.061

The main aim of the present study is not model comparison, but to test a reciprocal effects model, as well as to investigate the relationships between national test scores and level and facet-specific self-efficacy, respectively. Accordingly, parameter estimates from both models are reported (see Figures 19 and 20). As can be seen in Figure 19, results from Model 1 partially support a reciprocal effects model between levels of self-efficacy and national test scores (hypothesis 1). Specifically, national test scores in grade 8 significantly predict levels of self-efficacy in grade 9 in relation to medium (.19) and hard (.21), but not easy, tasks (hypothesis 2). The predicted relationship from self-efficacy level constructs in grade 8 to national test scores in grade 9 was not supported (hypothesis 3). However, all three levels of self-efficacy are associated with test scores in the same year, including when previous test performance are controlled in grade 9 (easy .15; medium .31; hard .22) (hypothesis 4). The results for hypothesis 3 and 4 might be related to hypothesis 5, which is based on the high

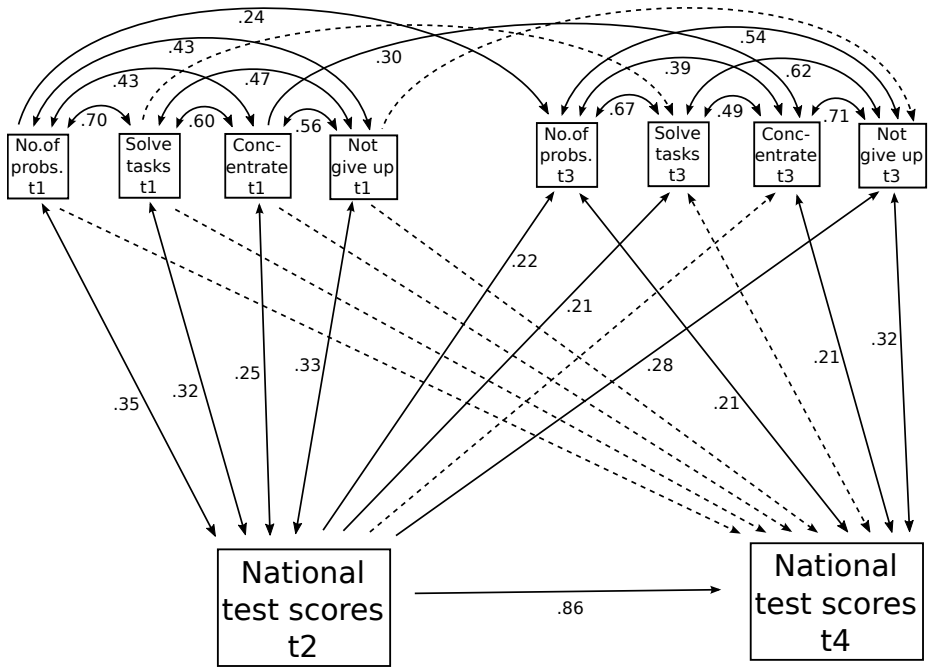
specificity of the self-efficacy measure. All items are explicitly linked with the immediately upcoming test, thus it is reasonable the self-efficacy constructs are associated with performance results of the immediately upcoming tests, but not with the tests one year later. Two further relationships are non-significant: the path from self-efficacy easy level in grade 8 to grade 9, and the path from national test scores in grade 8 to self-efficacy easy level in grade 9. It is possible that the easy level items were too easy ("ceiling effect") (see the discussion in section 3.4.1). However, the relationships with performance results are consistently stronger for self-efficacy medium and hard levels, than self-efficacy easy level (hypothesis 7). Finally, test scores are more stable from grade 8 to grade 9 (.87) than are levels of self-efficacy (.23 and .28 for medium and hard level tasks, respectively) (hypothesis 6).



Note: "Easy", "Med., and "Hard" refer to easy, medium and hard tasks. "t1" and "t2" refer to time points 1 and 2 (both in grade 8); "t3" and "t4" refer to time points 3 and 4 (both in grade 9).
All estimates are standardized (STYX) estimates from Mplus (Muthén and Muthén, 2012).
Paths are significant at the $p<.05$ level. Dotted paths are nonsignificant.

Figure 19: Structural equation model, self-efficacy level constructs (Model 1)

Model 2 results (see Figure 20) are generally consistent with results for Model 1. First, results for Model 2 including facet-specific self-efficacy are also partially in support of the reciprocal effects model (hypothesis 1). Test scores in grade 8 predict three facets of self-efficacy in grade 9 (not "concentrate") (hypothesis 2), while hypothesis 3 was not supported (the paths from self-efficacy constructs in grade 8 to test scores in grade 9). However, facet-specific self-efficacy constructs are related to performance results in each year (hypothesis 4), including in grade 9 (not "solve tasks") when previous test scores are included in the model. As discussed in relation to Model 1 results, the specificity of the self-efficacy items means it is reasonable to expect stronger associations with performance results in the same year, than one year later (hypothesis 5). Similar to the easy level constructs in Model 1, the proposed relationships between "solve tasks" and "not give up" from grade 8 to grade 9 are not significant.



Note: "No. of probs.", "Solve tasks", "Concentrate", and "Not give up" refer to facets of test taking.
"t1" and "t2" refer to time points 1 and 2 (both in grade 8); "t3" and "t4" refer to time points 3 and 4 (both in grade 9).
All estimates are standardized (STYX) estimates from Mplus (Muthén & Muthén 2012).
Paths are significant at the p<.05 level. Dotted paths are nonsignificant.

Figure 20: Structural equation model, self-efficacy facet constructs (Model 2)

Again as with Model 1 results, test scores are more stable over time than facet-specific self-efficacy (estimate is .86 for national test scores, while .24 and .30 for "number of problems" and "concentrate", respectively) (hypothesis 6). Unlike the other facets, the magnitude of the relationship between "not give up" and test scores is very similar in grade 8 and 9 (this magnitude is smaller in grade 9 for the other facets, when previous test results are included). The relationship between this facet and performance results is relatively consistent, in case of the relationship between scores each year, as well as from grade 8 performance scores to grade 9 self-efficacy scores. While the relationships between the other self-efficacy constructs and performance scores vary according to type of self-efficacy facet, there does not seem to be consistent patterns to these findings (hypothesis 8).

3.5 Discussion

In this study I investigated the theoretically proposed reciprocal effects model between level and facet-specific self-efficacy constructs, and performance results on national tests in mathematics as students progressed from grade 8 to grade 9. I used the established factor structure from a previous study (see study 1, presented in chapter 2) to specify two empirical models, where items were parcelled according to a level-structure and a facet-structure, respectively. Applying structural equation modelling, I used these two models to test for reciprocal effects, and to investigate differential relationships between the self-efficacy constructs and performance results.

Both models resulted in excellent fit, according to the selected criteria. Model 1, including parcels for self-efficacy levels, resulted in the best fit in absolute terms. The reason for the good fit is likely that Model 1 mirrors the factor structure obtained from the previous investigation of the self-efficacy measure (see study 1, presented in chapter 2). A model

where the variance from each test-taking facet is included in three different levels parcels (easy, med and hard) seems to most accurately represent the multidimensional nature of the measure. Also, the items associated with the easy level factor are (consistently) more non-normally distributed (see Table 24). Thus, partitioning these items in a single factor, rather than spreading them across every latent factor, likely leads to less biased parameter estimates (especially in relation to the medium and hard constructs), and better fit. I now discuss the findings, organized according to research question.

3.5.1 RQ1: Reciprocal effects

In both models, the reciprocal effects model was partially supported, in that self-efficacy were predicted by prior test scores, and vice versa. This is as expected, and in accordance with theoretical predictions from self-efficacy theory. However, there are differences between the relationships in regard to the strength of the associations, and the time elapsed between measurements. The relationships also differ according to the type of self-efficacy construct.

While national test scores generally predicted self-efficacy a whole year later, self-efficacy constructs in grade 8 did not predict national test scores in grade 9. However, grade 9 self-efficacy was significantly related to scores on the upcoming test in the same year. This is as predicted, given that the items presently used to measure self-efficacy are related to immediately upcoming tasks and challenges.

When comparing the effects between grade 8 and grade 9 constructs, the dominant effect is from national test scores predicting self-efficacy (the "skill development model"). The significant estimates range from .19 (medium level self-efficacy) to .28 (self-efficacy for perseverance), while none of the estimates for self-efficacy constructs in grade 8 predicting test scores one year later are significant. However, the correlations between self-efficacy

constructs in grade 9 and the same year test scores are either larger than (self-efficacy for medium level tasks at .31 versus .19) or of comparable magnitude (for instance .22 versus .21 in relation to hard level tasks) to the unique effects from grade 8 test scores on grade 9 self-efficacy constructs.

In my study, the relationship between self-efficacy and performance was evident in the short term (where the design of the study implies this relationship was directional), but long-term effects from self-efficacy to national test scores were not evident. It is likely that the effect of self-efficacy diminishes with the passing of time, in accordance with the recommendation that measures of self-efficacy are collected close in time to upcoming measures of performance (Pajares & Miller, 1995). In two previous studies, measures of self-efficacy and performance were collected over the course of a single day (Locke et al., 1984) or three years apart (Hannula et al., 2014). In both these studies, the findings indicated the dominant effect was from performance to self-efficacy. However, in the findings by Hannula et al. (2014), the difference in effect seemed to decrease with age. The type of test, and whether it is considered important by the students, might influence the relationship between performance experiences and self-efficacy. National tests are emphasised as important by Norwegian school authorities, and teachers are encouraged to use the results to discuss learning progress with the students. Also, students in grades 8 and 9 are likely becoming increasingly aware of upcoming exams (at the end of year 10). This might have provided quite a memorable context for their test experiences. If test performances have long-term effects on students' self-efficacy, this implies teachers should consider how students make sense of their mathematics test experiences, to support the development of adaptive self-efficacy.

3.5.2 RQ2: Differential relationships between self-efficacy levels and facets, and performance results

As well as differences in regard to time elapsed; the findings indicate the effects differ between types of self-efficacy construct. The strongest effects are between self-efficacy for medium and hard level tasks and their test scores, and self-efficacy for perseverance (not give up) and test scores. Assuming that students draw on level and facet specific experiences from the national test in grade 8 to form their level and facet specific self-efficacy in grade 9, these findings provide an indication of what level tasks (medium and hard) as well as what facet of test taking (perseverance) provided most salient mastery experiences during the national tests in grade 8. Furthermore, self-efficacy for medium level tasks and self-efficacy for test perseverance provided the highest estimates predicting scores on same-year national tests.

In the present study, both self-efficacy and national test scores are relatively stable, with a high portion of individual differences in grade 9 predicted by grade 8 scores, particularly in regard to test scores (.87 and .86 in models 1 and 2, respectively). Stability differs between the self-efficacy constructs. Self-efficacy beliefs for easy level tasks are not significantly associated across the two grades, and are also not predicted by previous test scores. In comparison, self-efficacy for medium and hard level tasks alike have both a stable (grade 9 self-efficacy scores significantly predicted by grade 8 self-efficacy scores) and a dynamic (grade 9 self-efficacy scores predicted by grade 8 national test scores) component.

Self-efficacy for test concentration is the most stable of the self-efficacy constructs, where grade 9 scores are predicted by the scores from grade 8 but not by previous test scores. In comparison, grade 9 self-efficacy for test perseverance and solving tasks of a specific challenge seem less stable, as they are not significantly predicted by grade 8 self-efficacy

scores, but are predicted by grade 8 national test scores. Finally, self-efficacy for solving a number of problems also have both a stable and dynamic component, predicted by grade 8 scores on the construct, as well as grade 8 test scores.

3.6 Conclusion

I used a recently developed measure of mathematics self-efficacy that includes levels and facets of self-efficacy, to investigate the theoretically proposed reciprocal relationship between level and facet-specific self-efficacy, and performance in mathematics (RQ1). In accordance with findings from the few previous investigations of reciprocal effects, it was demonstrated that students' mathematics self-efficacy were associated with scores on a test in the same year, while national test scores in mathematics predicted their self-efficacy one year later. This lends support to both the skill development as well as the self-enhancement models. Student mathematics self-efficacy might be supported through increased skill development, and positive self-efficacy might likewise support their learning and performance. Further studies should investigate the development and change of self-efficacy and their relationship with performance experiences across varying time spans, to increase our understanding of this dynamic relationship.

Furthermore (RQ2), differential relationships were found between national test scores and easy, medium and hard levels of perceived difficulty. In line with study 1, the associations with overall test scores were strongest for self-efficacy for medium levels of difficulty (the cross-lagged effect from national test scores in year one were comparable between medium difficulty and hard self-efficacy). The relationships with test performance also varied between different facets of test taking, where the strongest association was between perseverance and test scores. These findings indicate it may be useful for students to work on tasks of medium level difficulty, to develop accurate self-efficacy as well as skills in mathematics.

Furthermore, they indicate that as well as problem solving self-efficacy, self-efficacy for self-regulation is related to students' subsequent test performance. This implies classroom time might be well spent developing self-regulatory as well as problem solving skills.

The current findings are limited in that I could not control for other factors that may have influenced students' self-efficacy and test performance. Also, while the measure of self-efficacy provided interesting information, partitioning the effects of test taking facets and levels of perceived difficulty, further replications are needed with a larger sample, and in different cultural contexts. With exception to the sample, the further limitations discussed for study 1 (section 2.5.4) pertain also to this study.

4 Stability and change of students' self-efficacy when learning a new topic (study 3)

Self-efficacy in mathematics is important because it is related to engagement and persistence in mathematics, and students' choices to study and choose mathematics related careers. While theory posits self-efficacy is dynamic, and changes as an effect of continued appraisals of personal experiences, I could find no previous studies investigating how students' mathematics self-efficacy changes from lesson to lesson during regular classroom learning. Furthermore, while theory posits levels of difficulty are central to students' self-efficacy, we do not know whether and, if so, how changes in students' self-efficacy strength vary as an effect of perceived level of difficulty.

In the present study I aimed to address these gaps by investigating the stability or change of students' self-efficacy across a sequence of 3-4 lessons in mathematics. I followed 9 classes of Norwegian grade 6 (n = 81) and grade 10 students (n = 100) as they were introduced to a new topic. I used a newly developed multidimensional measure, adapted from phase 1 in the present thesis, and students reported their self-efficacy for easy, medium difficulty, and hard tasks at the beginning of every lesson. Autoregressive simplex models for three self-efficacy constructs demonstrated that the rank-order stability of students' self-efficacy was high, and that students' self-efficacy was influenced mainly through their most recent self-efficacy changing experience. Furthermore, latent growth models demonstrated that the mean-level of students' self-efficacy changed in a non-linear fashion, where the change was characterised by steep initial growth, followed by a flattening trend over time. There were significant differences between the growth of self-efficacy according to age as well as levels of difficulty, where steeper growth slopes were associated with younger students and harder tasks.

The findings highlight the role of cognitive appraisals as a likely stabilising factor. This implies teachers need to ascertain how students make sense of their experiences when learning mathematics, and work in the long term to encourage adaptive appraisals, in order to support positive self-efficacy. Furthermore, if changes to students' self-efficacy occur mainly early on when learning new topics, this emphasises the importance of how new topics are introduced, and how students' sense-making is supported.

4.1 Introduction

Numerous studies have provided empirical support to the theoretical notion that students' mathematics self-efficacy is important (Hackett & Betz, 1989; Klassen & Usher, 2010; Pajares, 1996b). Self-efficacy in mathematics is important because it predicts students' engagement, effort, end enjoyment of mathematics, and future engagement with mathematics through study and career choices. Self-efficacy theory predicts that holding somewhat

overconfident self-efficacy is beneficial to student learning and academic performance, because self-efficacy is related to adaptive learning behaviours (Bandura, 1997; Klassen, 2004). In this way, positive self-efficacy encourages students to make use of their full potential, rather than underperform due to doubts of their own capabilities.

Knowledge about how self-efficacy develops in broad terms across several months or years is important in order to better understand the relationship between self-efficacy and factors such as grade level, school changes, curriculum change, and increasing exam pressures. In a shorter time frame, knowledge about changes in self-efficacy is important in order to develop better understanding of the relationship between self-efficacy and more immediate factors such as classroom mastery experiences, task structures, and teacher feedback and/or interventions. As such, the micro-analytic perspective can provide valuable insight into the processes of student learning and changes in self-perception, while the broader perspective is one step removed from these processes, and is apt to provide insight into changes in learning and motivation as they relate to contextual factors interacting with the learning process.

In the present study, I focus on changes in self-efficacy in a shorter time frame. I argue it is valuable to investigate to what degree students' self-efficacy varies across lessons in mathematics. Knowledge about how students' self-efficacy develops while learning mathematics, is valuable for teachers in order to understand how best to support the development of adaptive self-efficacy, as well as to develop knowledge about how to "break" maladaptive circles of (too) low self-efficacy. This kind of knowledge can be useful to teachers when they plan their lessons, and when they consider the kind of support and feedback they make available to students during classroom learning of mathematics.

While previous research indicates a decline in several motivation constructs with student age (e.g. Caprara et al., 2008; Gottfried, Fleming, & Gottfried, 2001; Helle, Laakkonen, Tuijula, & Vermunt, 2013; Hornstra et al., 2013; Jacobs, Lanza, Osgood, Eccles, & Wigfield, 2002) particularly from the primary to the secondary stage (Midgley et al., 1989), we know little about how students' mathematics self-efficacy changes across classroom lessons. Self-efficacy is held to be a relatively dynamic construct (Bandura, 1997), presumed to vary across subjects or tasks, as well as between situational contexts. Several researchers (e.g. Hannula, 2012; Schmitz, 2006) have argued for an increased awareness of the state/trait dimension of motivational constructs, and empirical investigations to increase our knowledge of their variability. Many studies of self-efficacy have included only a few measurement points, e.g. one or two, generally across the course of a semester or a year. This limits the ability to investigate changes in self-efficacy over time. As such, while self-efficacy beliefs are theoretically held as relatively dynamic states, many studies of self-efficacy have treated it as a trait-like, e.g. mostly unchanging, construct. This has also limited the possibility to investigate to what degree and in what contexts self-efficacy is best considered states or traits. A further design consideration is the temporal distance between measurement points, in order to capture relevant changes. It is possible that general trends (e.g. increase or decline in self-efficacy across several years) are contradicted by fluctuations in self-efficacy across shorter time frames (e.g. fluctuations across and within learning and exam situations). Furthermore, the shape of self-efficacy changes might be linear, or they may be linked to specific events or processes (curvilinear: quadratic or cubic shape).

As I discussed in study 1 (see section 2.2.1) previous studies of self-efficacy have differed in the degree of specificity, but have generally not included different levels of difficulty. As of yet researchers have not ascertained the shape of self-efficacy change. It is plausible that the

variability of self-efficacy differs as a function of the degree of specificity of the construct (more general self-efficacy might be more stable than more specific self-efficacy), as well as the level of perceived difficulty of the task/event considered. As such, it is important to consider what the appropriate unit of analysis is, in order to investigate the process of change of self-efficacy.

In the present study, I have chosen to focus on lesson-specific self-efficacy including level, strength, and facets of self-efficacy. In this context, I have included items specifically related /linked with classroom lessons in mathematics. I used several measurement points, to enable a micro-analytic investigation into the stability versus change of mathematics self-efficacy across classroom lessons in mathematics. I collected measures of multidimensional self-efficacy from students in nine classes at the beginning of a sequence of 3-4 mathematics lessons, in order to investigate the variability of self-efficacy across lessons and whether self-efficacy variability differed as a function of levels of perceived difficulty. Having laid out the importance of research into changes of mathematics self-efficacy, in the following section I will review previous research on changes in self-efficacy, and how they have contributed to knowledge.

4.2 Theoretical background

According to self-efficacy theory, self-efficacy develops through a dynamic process, in a mutual relationship with performance. Theoretical perspectives as well as empirical evidence indicate the strongest source for changes in mathematics self-efficacy is performance experiences in mathematics, and students' interpretations of these (Butz & Usher, 2015; Byars-Winston et al., 2017; Usher & Pajares, 2008). Students' self-efficacy influences their performance / learning experiences, which in turn influence their self-efficacy (Bandura, 1997). Importantly, performance experiences do not influence self-efficacy directly, but are

cognitively appraised by the individual. In other words, students interpret their performance experiences in regard to a range of factors, such as conceptions of ability, perceived task difficulty, effort, the amount of help they received, and other factors such as memory of past events (Bandura, 1997, p. 81). Over time (through feedback loops, where students make appraisals of past performance events), the reciprocal influences between self-efficacy and performance trend toward *equilibrium* (see Williams & Williams, 2010), where over- or under-confidence gives way to more accurate self-efficacy (in relation to individuals' actual performance capability).

Bandura (1997) emphasised it is not the passing of time itself, but rather experiences of influential events, that lead to changes in self-efficacy. As such, what is a "reasonable" approach in regard to the number of measurement occasions and the distance between them need to be considered in relation to the aims and focus of each individual study. While a broader lens (e.g. measurement points across several years) can provide insight into self-efficacy changes in relation to e.g. student cognitive development and school, curriculum, and teacher changes, these factors are removed from the theoretically assumed external main drivers (sources) of the change process. In contrast, individual mastery experiences occur on a daily basis during lessons in mathematics. According to theory (e.g. Usher & Pajares, 2008), students make use of available information to repeatedly assess their experiences in mathematics, and to form their mathematics self-efficacy. Thus a micro-analytic focus is essential in order to understand the processes through which self-efficacy develops during classroom learning. In order to investigate change of a construct, a minimum of two time points are necessary. If the change is non-linear (non-monotonic change), at least three time points are needed. Schmitz and Wiese (2006) conceptualised learning as a process, and concentrated on sequences of (daily) learning states across five weeks to investigate the

process of self-regulated learning. Malmberg et al. (Malmberg, Pakarinen, Vasalampi, & Nurmi, 2015) included on average 11 "learning episodes" per student to investigate students' situation-specific motivation during one week. As my review of the literature will demonstrate, investigations into changes of self-efficacy within any time frame are relatively few. This may be related to the cost of longitudinal investigations, in regard to resources and time.

I am aware of only few studies that investigated self-efficacy change across several years, for instance from grade 3, to 6, to 9 (Hannula et al., 2014), or once a year from age 12 to 18 (Caprara et al., 2008). At the other end of the spectrum, on a smaller time scale, I am aware of only one study (Bernacki et al., 2015) that investigated changes in self-efficacy across learning tasks, e.g. every few minutes. Most commonly, changes in self-efficacy have been investigated over the course of a single year (Bong, 2005; Friedel et al., 2010; Johnson, Edwards, & Dai, 2014; Pajares & Graham, 1999; Phan, 2012). As the following section will demonstrate, the findings from these studies have been mixed, with effect sizes / autoregressive coefficients representing stability of self-efficacy ranging from -.05 to .37. I will now go into detail about how previous studies have contributed to our knowledge about changes in students' self-efficacy over time. While the focus of my investigation is change in mathematics self-efficacy, studies focusing on self-efficacy in other domains are included (given the lack of empirical investigations into this theme).

4.2.1 Changes across several years

Previous studies have indicated that students' motivation as a whole decline as they progress through the formal education system. This has been observed in regard to constructs such as intrinsic motivation, self-regulation, task-value, and task-orientation (e.g. Gottfried et al., 2001; Helle et al., 2013; Hornstra et al., 2013; Jacobs et al., 2002). Furthermore, research has

indicated that students' perceptions of own competence decline with age (Benenson & Dweck, 1986; Jacobs et al., 2002; Nicholls, 1978; Stipek, 1981). At the same time, studies have found an increase in the rank-order stability of motivational constructs with age, e.g. the relative standing of individuals within a group (Gottfried et al., 2001; Wigfield et al., 1997). In regard to self-efficacy, Schunk & Pajares (2010) (in a non-empirical chapter) argued students' self-efficacy starts declining around middle school. However, the long-term development of students' self-efficacy has rarely been investigated within a single study. Rather, it appears findings from different cross-sectional studies (or samples) are relied upon to make this assertion. Zimmerman and Martinez-Pons (1990) included the mathematics self-efficacy of 5th, 8th and 11th grade students as dependent variables and investigated whether there were statistically significant differences between the means of the three samples. Contrary to the assertion by Schunk and Pajares (2010), Zimmerman and Martinez-Pons (1990) demonstrated that grade 8 students' mathematics self-efficacy beliefs were significantly *higher* than those of students in grade 5.

One example of a longitudinal study of self-efficacy is the one by Caprara et al. (2008), who used a staggered, multiple cohort design to follow 412 Italian junior high school students from age 12 till age 18 (grade 6 to 12). The authors demonstrated students' self-regulatory self-efficacy (example item *How well can you organize your school work?*) did decline over time in broad terms. Students' self-efficacy declined in a linear fashion at β =-.08 for boys and β =-.04 for girls, every two years (a nonlinear model fit the data equally well). Inspecting the correlation matrix, the authors demonstrated that students' self-efficacy was relatively stable over time (most correlation coefficients were between .43 and .70), where the strength of the associations decreased with longer distance in time between the measures, demonstrating a simplex structure of association (Little, 2013). The authors used repeated

measures of self-efficacy, however they did not comment on whether measurement (factorial) invariance was established. Caprara et al. (2008, p. 530) argued that a decline in students' self-regulatory self-efficacy might be related to increasing academic demands in higher grades, where cumulative academic "deficiencies" become more salient, affecting their self-efficacy. Furthermore, they argued a decline in students' self-efficacy might be related to a decline in teachers' efficacy to positively influence students' motivation and learning in the older grade levels. Finally, they argued students' self-efficacy decline might be related to an increase in competing activities and interests.

Similar to Caprara et al. (2008), Hornstra, van der Veen, and Peetsma (2013) investigated self-efficacy changes across several years. The authors focused on students' (N= 722) academic self-efficacy across five waves, from grade 3 (age 9) to grade 6. The authors used latent growth curve modelling, and found that a quadratic model best described the changes in students' self-efficacy over time. Students' self-efficacy initially declined, before increasing, leading to no overall decline in self-efficacy. The authors commented that previous mixed findings on changes in students' self-efficacy over time may be due to the curvilinear shape of self-efficacy change.

Another couple of studies collected students' self-efficacy across several years, however the focus of these studies was not self-efficacy change. Nevertheless, the studies reported findings regarding the stability of self-efficacy over time. Caprara et al. (2011) investigated the relationship between academic self-efficacy (e.g. across several subjects) and academic performance. They used the same sample as in the study by Caprara et al. (2008), but focused on students as they progressed from age 13 to age 19 (grade 7 to 13). Academic self-efficacy was measured at ages 13 and 16, with 15 items including two domains of self-efficacy: to

master different subject/curricular areas (example item *How well do you do in mathematics?*), and to self-regulate their own learning (example item *How well can you study when there are other interesting things to do?*). The authors used exploratory factor analyses to determine their self-efficacy measure was unidimensional, however they did not report on measurement invariance (beyond configural invariance). Students' mean self-efficacy scores declined in absolute terms for boys (from 3.78 to 3.60) while they remained practically unchanged for girls (from 3.74 to 3.76). The differences were not tested for statistical significance. Looking at the correlation matrix, self-efficacy was "significantly" stable over time (r= .50/.45, boys and girls respectively). The authors specified autoregressive effects in a structural equation modeling framework, and found that the stability parameter for self-efficacy was β = .30 /.31 (boys/girls) from 7th to 10th grade.

Similarly to Caprara et al. (2011), Hannula et al. (2014) reported students' mathematics self-efficacy remained relatively stable across time, as students progressed from grade 3, to grade 6, to grade 9. As a measure of self-efficacy the authors used a four-item "mathematics self-efficacy scale" (example item *Mathematics is an easy subject*). It should be noted the example item provided differs in face validity to items commonly used to measure self-efficacy (Bandura, 2006; Klassen & Usher, 2010). Rather than a measure of students' confidence they *can do* mathematics, it is an evaluative statement regarding whether or not they perceive mathematics *to be an easy subject*. It is conceivable, for instance, that students may find mathematics hard relative to other subjects and, yet, have high confidence in their ability to do mathematics. The authors reported autoregressive stability effects of β = .36 (grade 3 to grade 6) and β = .49 (grade 6 to grade 9). In other words, the rank-order of self-efficacy at grades 3 and 6 predicted the rank-order of self-efficacy at grades 6 and 9, respectively. In the study by Hannula et al. (2014), students' self-efficacy increased in stability over time. The

findings by Hannula et al. (2014) thus indicate it might be easier to influence younger, than older, students' self-efficacy.

4.2.2 Changes across a single year

The studies by Caprara et al. (2011) and Hannula et al. (2014) provided evidence of the relative stability of self-efficacy across several years. Findings on changes in self-efficacy over shorter time spans have been mixed. Bong (2005) measured students' self-efficacy at four time points, before and after two examinations. She found that within a single year (April to February) 375 South Korean freshmore (e.g. grade 10 US) high school girls' mathematics self-efficacy increased significantly, however there were also significant fluctuations in students' self-efficacy around exam times. Self-efficacy increased after the first examinations, decreased before the second examinations, and then increased again after the second examinations. The structural equation modelling regression paths of self-efficacy from one time point to another were: t1-t2 $\beta = .35$, t2-t3 $\beta = .37$, t3-t4 $\beta = .29$. To measure self-efficacy, Bong (2005) adapted items from the Self-Efficacy subscale of the Motivated Strategies for Learning Questionnaire (MSLQ; example item *I am sure that I can do an excellent job on the problems and tasks assigned for [schoolwork/ subject class]*; Pintrich & De Groot, 1990) and the Patterns of Adaptive Learning Scales (PALS; example item *I'm certain I can master the skills taught in [school/subject] this year*; Midgley et al., 2000). The author did not comment on measurement invariance. Unlike the other subjects included in the study (Korean and English), students' mathematics self-efficacy (as measured at each of the four time points) failed to predict their subsequent exam scores. This may be related to the fact the measure of self-efficacy was relatively unspecific, and not well targeted towards the performance measure. Bong (2005) used students' end of semester exam scores in mathematics as achievement scores, while the self-efficacy items were "anchored" to the subject, and did not mention the exam(s), nor tap any of the specific exam content. Bong (2005) pointed to social

cognitive theory (Bandura, 1997) to explain the fluctuations in students' self-efficacy. She argued students' self-efficacy beliefs were raised after sitting exams as an effect of mastery experiences, while they decreased prior to exams, as the testing occasions became more "imminent".

Friedel et al. (2010) investigated changes in mathematics self-efficacy as students (N=929) transitioned from 6th grade (elementary school) to 7th grade (middle school) in the US. Using repeated measures ANOVAs Friedel et al. (2010) found that students' mean self-efficacy did not change significantly across four time points (fall and spring in grade 6, and fall and spring in grade 7). However, students' self-efficacy did change as a function of student mastery goal orientation and perceived teacher mastery goal emphasis (e.g. "perceptions of the teacher's emphasis on understanding and exploring math concepts, finding new ways to solve math problems, rewarding effort, and learning from mistakes", p. 106). Namely, self-efficacy increased for students who reported an increase in mastery goal orientation (β = .10) or for students who went from a low to high perceived mastery goal emphasis teacher (β = .05), and decreased for students who reported a decrease in mastery goal orientation (β = -.11) or for students who went from a high to low perceived mastery goal emphasis teacher (β = -.05). Similar to Bong (2005), the authors used items (in this case five) from the Patterns of Adaptive Learning Survey (PALS; Midgley et al., 1997) to assess students' self-efficacy for learning and doing mathematics. The authors argued students' self-efficacy can be seen as indicators of psychological adjustment at times of transition (e.g. change of school environment), and are "sensitive to perceived changes in classroom goal structures" (Friedel et al., 2010, p. 110). Their findings highlighted the importance of not only change in perceived teacher goal emphasis, but of the quality and magnitude of this change, on students' self-efficacy.

Pajares and Graham (1999) investigated changes in task-specific mathematics self-efficacy (students' confidence they could solve specific items correctly) from the beginning (October) to the end (April) of the school year for 273 middle school (grade 6) students in the US. The authors used dependent samples t-tests to ascertain that self-efficacy as well as performance mean scores decreased significantly from the first to the second exam, Δ = -.79 and -3.29, respectively. The authors also calculated students' mean bias and accuracy scores, by contrasting their self-efficacy scores with their performance scores. They found that students' self-efficacy became more over-confident (Δ = .36), and less accurate (Δ = -.92) during the course of the year. Although students were more over-confident in spring than fall, higher self-efficacy predicted higher exam scores in a similar magnitude in spring (β = .27), as compared with the fall scores (β = .27). In other words, increases in mean over-confidence was not associated with a lower self-efficacy - performance relationship. To measure self-efficacy students were asked to rate their confidence they could solve each of 20 items. The items displayed to the students were similar to items in the upcoming test (example item *A train is traveling an average speed of 75 miles per hour. Use the four-step plan to find out how far it will travel in four hours*). The authors argued that the spring (end of the year) exams were more difficult than the fall exams, thus explaining the decrease in both self-efficacy and performance scores. However they concluded that more studies were needed (with performance events of equal difficulty) to ascertain whether their findings were a result of the increasing level of difficulty of the spring versus fall exam, or if their findings reflected a tendency that 6th graders' strength and accuracy of self-efficacy decreased throughout the year.

Phan (2012) used latent growth modelling to investigate the relationship between sources of self-efficacy, and changes in task specific mathematics self-efficacy across four occasions within one year. The participants were 3rd and 4th grade students in three elementary schools in Australia (N=339). Phan (2012) reported significant intercept (5.43) and slope (.34) factors (both unstandardised), indicating that students' mathematics self-efficacy increased across the four time points in a non-linear fashion (the author did not specify the shape of the change, e.g. quadratic or cubic, but rather allowed the slope parameters of the second and third measurement occasion be freely estimated from the data. This is equivalent to the *unconditional* growth models I investigate in the present study, in section 4.4.4.2). Phan (2012) used the 12 task-specific items from the ASQ (example item *How confident are you that you can solve: 45 + = 30?*), to measure students' mathematics self-efficacy. Phan argued his results indicated self-efficacy beliefs "fluctuate in an unpredictable manner" (2012, p. 203). Furthermore, he argued these fluctuations were likely linked to social factors, processes, and contexts, offering teachers' pedagogical strategies as one example.

Johnson, Edwards, and Dai (2014) also investigated changes in self-efficacy, however focusing on undergraduate educational psychology students (N=63). Students' self-efficacy was assessed at five measurement occasions across a 15 week semester, using eight items from the Motivated Strategies for Learning Questionnaire (MSLQ) (example item: *I'm certain I can understand the most difficult material presented in the readings for this course*). The authors used the sum scores of students' self-efficacy as manifest variables in a Latent Growth Model framework, and found that a cubic growth model fit the data best (slope mean = 3.9, quadratic mean= -4.24, and cubic mean= 0.78, unstandardised estimates). Their findings indicated students' self-efficacy declined after completing their first assignment, and rebounded toward the initial self-efficacy levels prior to the final exam. Students' self-

efficacy scores were collected prior to course assignments and exams, and the authors argued that the fluctuations in self-efficacy reflected how students constantly re-evaluate their capabilities after performance experiences. According to the authors, it is likely the students' confidence increased and they became more certain about what the final exam entailed, after receiving feedback from their previous exam and assignments. This mirrors the explanations offered by Bong (2005) and Pajares and Graham (1999); that students' self-efficacy fluctuates in relation to exam experiences. Importantly, Johnson et al. (2014) argued motivational decline is not inevitable, which stands in contrast with some literature referring to a general decline in motivation and self-competence constructs after middle school (e.g. Gottfried et al., 2001; Helle et al., 2013; Jacobs et al., 2002; Schunk & Pajares, 2010; Wigfield et al., 1997).

4.2.3 Changes across learning events

In a smaller time-frame still, Bernacki et al. (2015) investigated 9th grade students' (N=107) task–specific self-efficacy and problem solving performance while working on algebra problems within an intelligent tutoring system ("Cognitive Tutor Algebra"). Sequential prompts were embedded in the program to elicit self-efficacy from the participants after receiving (immediate and automated) feedback on their performance during learning. A single self-efficacy item (example item *How confident are you that you could solve a math question like this one in the future?*) was used to measure students' self-efficacy after the completion of every fourth algebra problem in the unit. Bernacki et al. (2015) referred to Ainley and Patrick (2006; as referenced by Bernacki et al., 2015) that a single item approach is valid as long as the construct is "unambiguous to the respondent". Bernacki et al. (2015) argued their item met this requirement. A total of four self-efficacy prompts were given across the learning unit, but the duration of the unit varied for individuals, as they progressed at their own pace. Typically, the students responded to a self-efficacy prompt every five minutes (personal communication with the first author, April 25th, 2017). Inspecting the correlation matrix, the

authors found that self-efficacy was quite stable across time, as all self-efficacy items were significantly correlated, with estimates ranging from $r = .61$ to $r = .86$. Indeed, at the group level, self-efficacy did not change between the time points. However, the authors found that within a single learning task, self-efficacy "varied reliably" at least once for 60% of the students. This indicated that some students experienced increases in self-efficacy and other students experienced decreases in self-efficacy, resulting in no overall mean change. The authors argued their findings provided evidence that self-efficacy does fluctuate during learning, and that students assessed the accuracy and speed with which they had solved previous problems, to formulate their self-efficacy.

4.2.4 Summary of research on self-efficacy change

The above studies have varied according to conceptualisation and instrumentalisation of self-efficacy (e.g. the "mathematics self-efficacy scale" by Hannula et al., 2014, versus measures congruent with social cognitive theory), facet-specificity (e.g. academic versus task-specific self-efficacy), and other important design choices (e.g. sample, number of measurement points, analytical procedures). While methodological differences allow for limited comparison across studies, the studies have nevertheless highlighted some broad trends (common across most of the studies).

First and foremost, I can find no empirical evidence of a general decline in mathematics self-efficacy, with age /grade level. Only two of the studies I reviewed reported a decline in self-efficacy: however, the study across several years by Caprara et al. (2008) was concerned with self-efficacy for self-regulation, and Pajares and Graham (1999) argued the decline in mathematics self-efficacy they found across one school year likely was due to differing contextual exam pressures at the two measurement points they used. The other studies reported positive change, no overall change, or fluctuations in students' self-efficacy, across

different time spans (see Figure 21 for a schematic overview of empirical findings from the studies I reviewed).

Second, correlation coefficients and autoregressive path estimates from the studies reviewed indicate there is significant stability in self-efficacy across time, meaning the rank-order of students in terms of their self-efficacy remains relatively stable over time. However, stability of the standing within a group (over a length of time) could be coupled with significant changes in mean levels for the whole group, meaning there may be high stability of the rank-order of students, while also substantial changes to the mean level of their self-efficacy scores over time.

To this effect, several of the studies I reviewed provide empirical evidence students' self-efficacy fluctuate in relation to individual and contextual factors, such as perceptions of goal emphasis or exams. The study by Bernacki et al. (2015), focusing on problem solving within an automated algebra tutor, indicated there is significant variation in students' self-efficacy across problem solving tasks, even when group-level changes are not evident. Accordingly, it seems likely there is significant self-efficacy variability at the individual level and across shorter time frames, also in studies that highlighted the overall stability of self-efficacy (e.g. Hannula et al, 2014, Caprara 2011).

As we can see in Figure 21, three of the studies investigated changes across several years, five studies investigated self-efficacy changes within one year (e.g. across a semester or a change of schools), and only one study employed a micro-analytic approach, investigating self-efficacy during learning within an automatic algebra tutor. Furthermore, three of the studies (Bernacki et al., 2015, Pajares & Graham, 1999, and Phan, 2012) included task-specific

measures, four of the studies focused on subject level mathematics self-efficacy (Bong, 2005, Friedel et al., 2010, Hannula et al., 2014, and Johnson et al., 2014), while the two studies by Caprara and colleagues (2011 and 2008) investigated academic self-efficacy and self-efficacy for self-regulation, respectively.

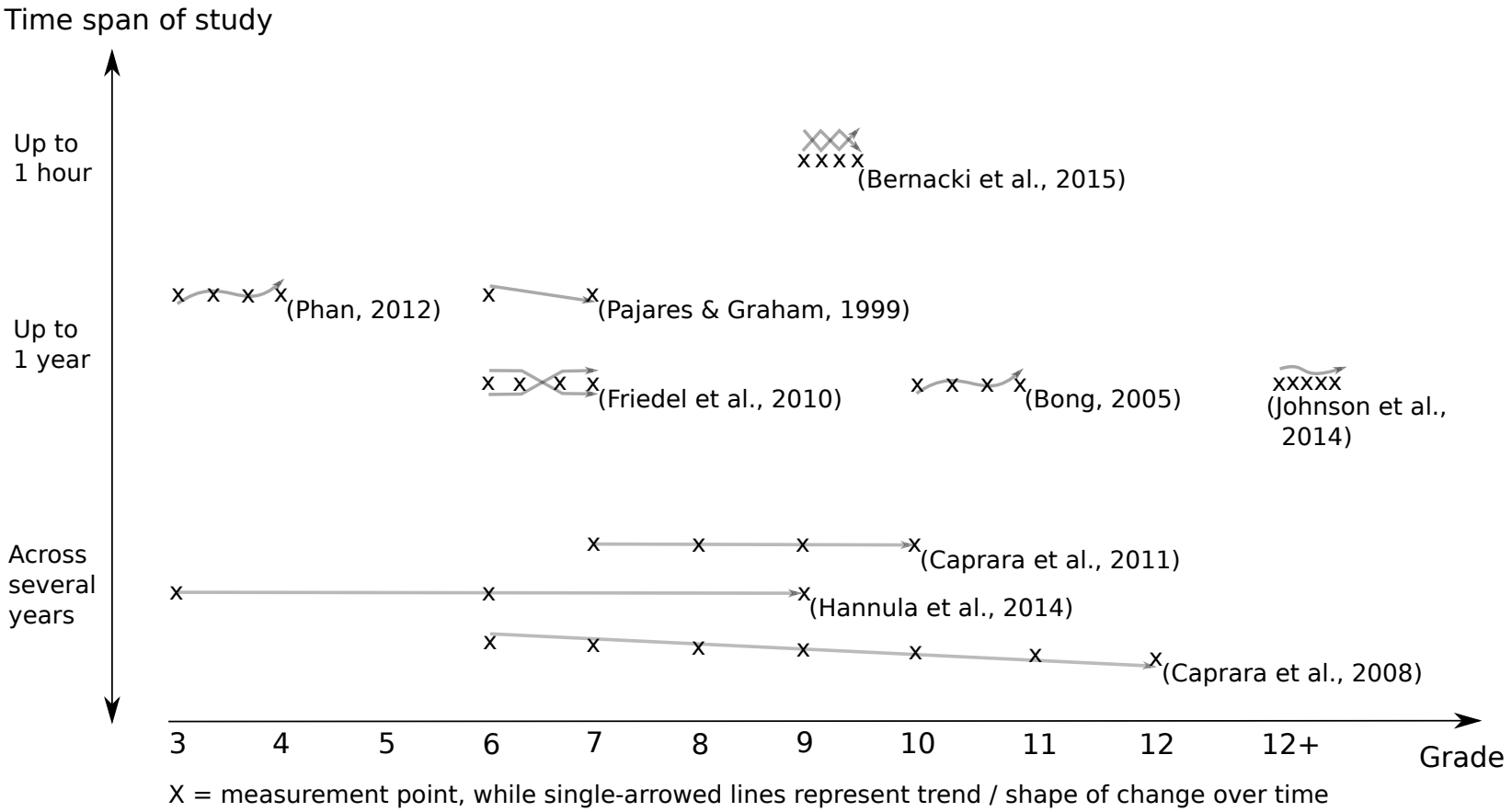


Figure 21: Overview of empirical findings on Self-Efficacy change

From a theoretical point, I would expect that self-efficacy measured at the most specific levels (e.g. task and problem-specific self-efficacy) change to a larger degree as compared with self-efficacy measured at more general levels, such as mathematics or academic self-efficacy. From the empirical studies I reviewed, we cannot conclude whether the variability of self-efficacy is indeed related to the facet-specificity of the self-efficacy measure itself.

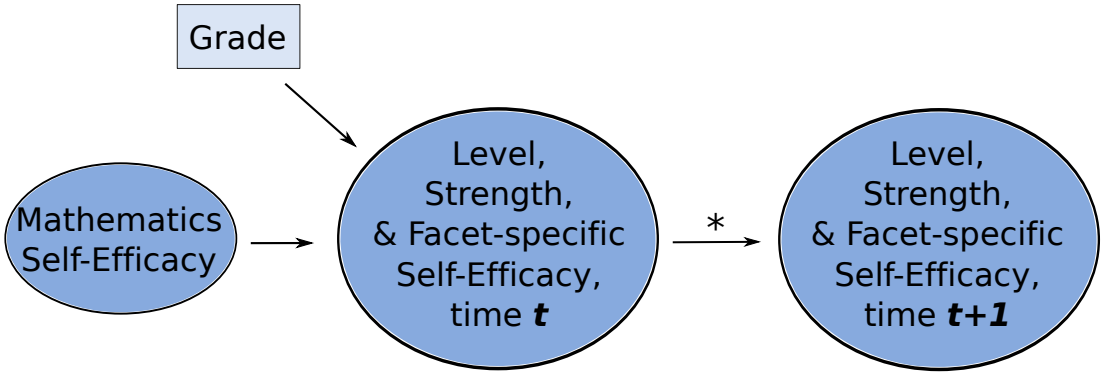
As mentioned, it is hard to make broad or definite conclusions about the manner in which self-efficacy changes, given that we have to dis-attenuate differences in results related to sample or context (e.g. facet specificity), from differences in methodological approaches (e.g. number and temporal distance of measurement points).

Although some authors refer to a "progressive decline" in students' motivation constructs as they go through the school system (e.g. Caprara, 2008), I have not found any empirical studies that present evidence to this effect, in relation to students' mathematics self-efficacy. I argue we cannot conclude whether there are overall trends of increase or decrease in students' mathematics self-efficacy in the long term. In the short term, only one study investigated changes in self-efficacy in a micro-analytic approach, during student learning of mathematics. However, this study focused on learning within an intelligent tutoring system, rather than regular classroom learning. Furthermore, the study by Bernacki et al. (2015) used a single self-efficacy item, meaning it is hard to dis-attenuate measurement error from "actual" change in the construct over time.

4.2.5 Theoretical model and research questions

On the basis of theory and previous studies, I propose a theoretical model of self-efficacy change over time (see Figure 22). In this model students' more general mathematics self-efficacy predicts their self-efficacy for learning mathematics during lessons. Furthermore,

grade predicts self-efficacy at the beginning of lesson 1 negatively, according to decreasing competence beliefs with age (e.g. Jacobs et al., 2002). There is significant stability of rank-order over time, however self-efficacy is likely to vary at the individual level according to influential events such as mastery experiences, and self-efficacy mean might change across time at the group level.



* The arrow represents change over time, either in terms of rank order stability, or in terms of mean-level changes
Mathematics Self-Efficacy = Students' subject specific self-efficacy, measured prior to the first lesson in mathematics

Figure 22: Theoretical model for study 3

While some previous studies have collected data on self-efficacy at several time points, sufficient to analyse variation in the construct, I could find no studies that investigated changes in self-efficacy across regular classroom lessons. While relatively few studies with quantitative data have focused on changes in self-efficacy at all (as compared with the extensive literature concerning cross-sectional investigations), the lack of investigations into the change of mathematics self-efficacy across regular lessons in mathematics is a pronounced gap in the literature. Furthermore, as discussed in study 1, previous studies of mathematics self-efficacy have failed to include levels of difficulty as an explicit dimension

in their measures. Thus, we do not know whether changes in students' self-efficacy differ as a function of the levels of perceived difficulty of the task/event in question.

In order to address the above gaps in the research knowledge, I propose the following research questions:

(1) *What is the structural validity of the proposed self-efficacy constructs, which include facets and levels of self-efficacy?*
(2) *How stable is the rank-order of students' mathematics self-efficacy across a sequence of classroom lessons in mathematics?*
(3) *How do students' mathematics self-efficacy change across a sequence of classroom lessons in mathematics?*

Having presented my research questions, I will now discuss how to best investigate these questions empirically.

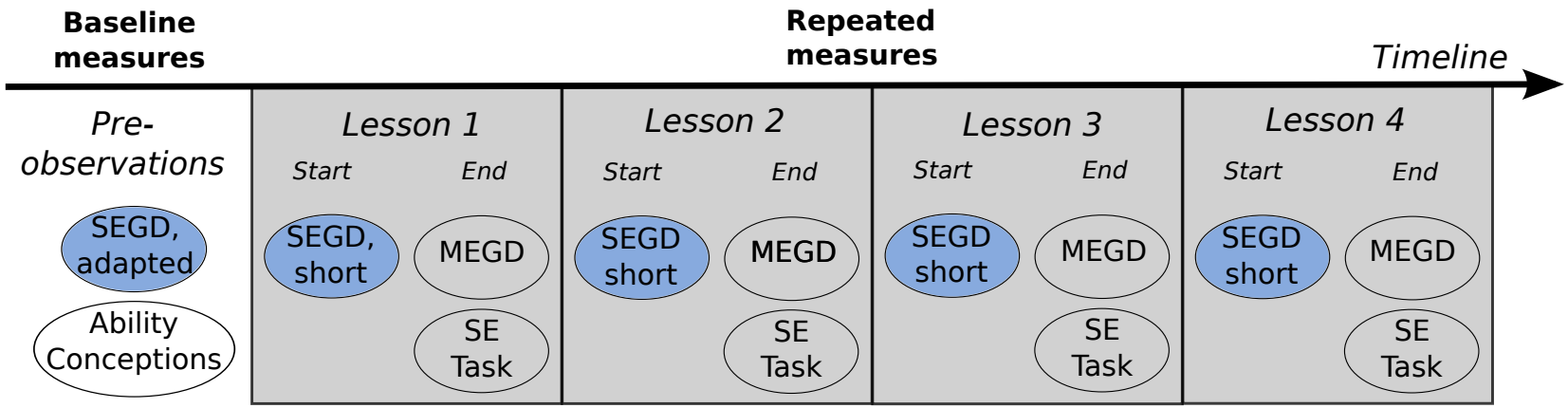
4.3 Methodology

The Classroom Phase (phase 2 of the thesis) was designed so as to build on and extend the findings from The National Test Phase (phase 1 of the thesis). While the primary aim in study 3 was to construct substantive knowledge through investigating changes in self-efficacy across classroom lessons in mathematics (research questions 2 and 3 in the present study), a secondary aim of this study was to investigate the validity of a multidimensional approach to measuring self-efficacy across lessons in mathematics (research question 1 in the present study), using a short version of the SEGD investigated in study 1. In this chapter I will describe the methods pertaining to the current study of my thesis. Where relevant I will refer back to study 1, in order not to repeat too much information. However, all aspects I consider central to this study will be addressed here to some degree, thus I might "err on the side of repetition" rather than not being detailed enough. I will first describe the overall design of the

phase two data collection, before discussing details about participant selection, measures, and analytic strategy.

4.3.1 Overall design

Phase 2 includes intensive longitudinal data, in order to illuminate the change in students' self-efficacy beliefs over the course of a sequence of lessons in mathematics (approximately one week per class). Students in four primary and five secondary school classes in Norway were included (N=181). Two researchers (myself and a research assistant, Ingrid Syse) followed each class for 3-4 lessons in mathematics when the teacher introduced a new topic. I wished to follow each class when they were introduced to a new mathematical topic because previous research has indicated self-efficacy might fluctuate according to external events (e.g. around exam times, see Bong, 2005; Pajares & Graham, 1999), and are most likely to change when students learn new things or engage with new experiences (see Usher & Pajares, 2008). Thus, the timing of the data collection was strategic rather than random. Students filled out questionnaires prior to the first lesson in mathematics, as well as short questionnaires at the beginning and end of each lesson. The end-of-lesson questionnaires are not included in the present study, but will be investigated in study 4. An overview of the design of phase 2 is presented in Figure 23.



NB! Measures in blue ovals were used in study 3; measures in ovals, without fill, were included in study 4
SEGD = Self-Efficacy Gradations of Difficulty (adapted and short version);
MEGD = Mastery Experiences Gradations of Difficulty; SE Task = Task Self-Efficacy

Figure 23: Design of phase 2

The types of measures collected (in ovals) are organised along a horizontal timeline for the data collection. To the left is listed data that was collected prior to the first lesson of each class (the pre-observation time point), while in the grey boxes to the right data is listed what was collected during each consecutive lessons in mathematics. Overall, my design includes five measurement occasions: pre-observation (time point 0), followed by lessons 1 to 4 (time points 1 – 4). It is important to note the distance in time between each mathematics lesson /measurement occasion differed between the classes, which has implications for the interpretations of my findings (see section 4.3.6.2).

The measures I used for investigations in study 3 are in blue ovals (Figure 23). In order to capture students' baseline mathematics self-efficacy, I used an adapted version of the Self-Efficacy Gradations of Challenge (SEGD, classroom version) questionnaire (Appendix J), originally developed for study 1 of The National Test Phase of the present thesis. To measure subsequent changes in students' self-efficacy through repeated measures, I used a short version of the Self-Efficacy Gradations of Difficulty (SEGD, short classroom version) questionnaire, adapted for the present study (Appendix K). The measures and their development are described in more detail in section 4.3.5.

4.3.2 Participants

Participants were Norwegian students in 6th grade (4 classes, n= 81) and 10th grade (5 classes, n=100), approximately 11 and 15 years old, respectively. A large percentage of students in each class consented to participate (89% overall) after receiving written information about the study (see section 4.3.3). Numbers of students and accompanying missing data varied somewhat from lesson to lesson, mainly due to absenteeism (illness). It was also common that one or two students in each class did not take part in the regular mathematics class. For instance, several non-native speakers of Norwegian took some form of adapted math classes

(in Norwegian: "spesialundervisning"). Table 27 presents the number of participants from each class at each time point for the study.

Table 27: Phase 2 participants at each time point

Participants (girls)		Pre-obs.	Lesson 1	Lesson 2	Lesson 3	Lesson 4
Grade 6	***81 (44)***					
Class 21	*25 (11)*	24	24	23	21	21
Class 23	*23 (12)*	23	23	20	21	22
Class 24	*23 (14)*	18	21	21	21	14
Class 31*	*10 (7)*	10	10	*MBD*	9	*MBD*
Grade 10	***100 (46)***					
Class 10	*25 (11)*	24	19	20	18	19
Class 11	*19 (7)*	19	16	15	16	15
Class 20	*24 (12)*	23	21	20	22	22
Class 35	*16 (8)*	14	16	16	15	*MBD*
Class 38	*16 (8)*	16	16	16	16	16

*I followed class 31 during two double lessons, coded as lessons 1 and 3, respectively.
MBD = missing by design

I used a strategic sample, where I requested to follow each class for as long as they worked on a new topic in mathematics. Data at the student level was collected at three to five time points per student, depending on whether the teacher focused on the new topic over three or four lessons. Given that the number of lessons varied between the classes, some data in my study is missing by design (see Table 27). As an example, one grade six class was followed for two double lessons, while most other classes were followed for four single lessons. The classes varied in regard to which topic they focused on during the sequence of lessons. Furthermore, for each class, the lessons varied in regard to the main task the teacher planned for each lesson. In Table 28 I have listed the topics and tasks for each class / lesson. It is plausible that variability in self-efficacy changes is related to the classroom tasks students engage with. While interesting, that theme is beyond the scope of the present thesis, and will be investigated in a future study. In section 4.4.4.1 I present the intra-class correlations (ICCs) as an indication of the proportion of variance at the class versus the student level.

Table 28: Topic and main tasks for each class / lesson

Grade 6	Topic	Lesson 1 task	Lesson 2 task	Lesson 3 task	Lesson 4 task
Class 21	*Fractions*	*to split figures together with your "learning partner"*	*to draw figures so they become a whole (one)*	*to compare fractions*	*to find equal fractions in the fractions-game*
Class 23	*Fractions*	*to fold figures in order to understand fractions*	*to solve tasks regarding fractions from the book*	*to solve tasks regarding fractions and decimal numbers from the book*	*to solve tasks regarding fractions and decimal numbers*
Class 24	*Fractions*	*to solve tasks including fractions from the book*	*to solve tasks on fractions you went through during the lesson*	*to calculate from a part to a whole*	*to determine which fraction is greater or smaller*
Class 31	*Measurement*	*working in pairs to find different things to weigh, and to use the units kg, hg, and g.*	*MBD*	*working in pairs to solve tasks regarding scale-units*	*MBD*
Grade 10					
Class 10	*Equations*	*to solve tasks including equations*	*to solve equations graphically*	*to solve equations graphically*	*to solve equations with two unknowns*
Class 11	*Functions and equations*	*to solve tasks related to proportions*	*to solve tasks related to proportions*	*to enter functions in Geogebra*	*to enter functions in Geogebra*
Class 20	*Equations with two unknowns*	*to solve equations with two unknowns graphically*	*to solve equations with two unknowns using the substitution method*	*to solve equations with two unknowns using the substitution method*	*(to solve) tasks ("nuts") regarding finding the weight of skrews and nuts*
Class 35	*Geometry*	*to explore 90 degree triangles in order to understand Pythagoras'*	*to prove Pythagoras' theorem using Geogebra*	*to investigate triangles in order to determine whether they were 90 degree angled*	*MBD*
Class 38	*Geometry*	*to explain geometrical figures*	*(to solve tasks related to) Pytagoras` theorem*	*to use Geogebra in order to solve equations graphically*	*(to solve tasks related to) Pytagoras` theorem*

MBD = Missing by design

4.3.2.1 Selection

Phase 2 classes were selected from a pool of participants from phase 1, The National Test Phase. Phase 1 participants were selected through a joint selection strategy in a related project, The Learning Regions Project (see section 2.3.1 or Langfeldt, 2015; Sørlie & Söderlund, 2015). This project had aims to include schools from four Norwegian regions and with varying results on national tests, given known measures of socio-economic status in the municipalities. During phase 2 I aimed to include a range of classes, in regard to students' prior mean scores on national tests and self-efficacy in mathematics. These data were available from phase 1. Tables 29 and 30 present the prior mean scores on national tests in mathematics (absolute scores of 45 possible for grade 5, and 58 possible for grade 8 / 9), self-efficacy (a composite score of self-efficacy was used), and correlation estimates (for the test score / self-efficacy relationship) for the grade 5 classes that participated in 2013, and all grade 8 and grade 9 classes from phase 1 (in 2012 and 2013). Classes that participated during phase 2 are highlighted in yellow.

Some additional criteria guided my selection of classes: Due to aims of the related Learning Regions Project I selected classes from within two of the four regions (Sogn & Fjordane and Aust Agder, regions 1 & 2 in Tables 29 and 30). Furthermore, where possible I wished to include grade 10 classes that had participated since 2012 (as well as in 2013). The reason for this was the possibility to investigate longer trajectories of student self-efficacy change. Due to space limitations this theme is not included in the present DPhil thesis.

Table 29: Phase 1 test-scores, self-efficacy, and correlations, grade 5 classes

		Grade 5, 2013		
Region	Class	Test score	Self-efficacy	Correlation
1	**21**	29.88	7.83	.60**
3	22	27.13	7.66	.51
2	**23**	29.09	8.48	.20
1	**24**	30.38	7.96	.66**
4	25	28.67	8.54	.62*
3	26	31.83	7.74	.38
2	27	28.00	8.66	.09
2	28	24.38	7.03	.37
4	29	23.68	8.16	.74**
1	30	28.70	6.65	.28
1	**31**	20.60	9.00	.55*

	Test score	Self-efficacy	Correlation
Mean non-selected classes:	27.40	7.84	.37**
Mean selected classes:	27.99	8.24	.31**
Difference score:	0.58	0.41	
t-test:	0.48 (n.s.)	1.73 (p<.09)	

Legend: Class selected for phase 2

Table 30: Phase 1 test-scores, self-efficacy, and correlations, grade 8 & 9 classes

		Grade 8, 2012			Grade 9, 2013		
Region	Class	Test score	Self-efficacy	Correlation	Test score	Self-efficacy	Correlation
1	**10**	38.88	8.36	.16	38.75	8.01	.26
2	**11**	27.95	7.83	.64**	31.24	7.54	.46*
2	12	22.39	7.96	.33			
4	13	26.24	8.14	.42*	32.00	8.31	.46*
3	14	27.61	8.24	.39			
3	15	21.10	7.70	.31			
3	16	29.00	8.30	.34			
2	17	30.54	8.36	.72**			
2	18	41.50	8.21	.41			
4	19	30.60	8.43	.13	33.00	8.36	.73**
1	**20**	30.63	8.25	.46	36.67	8.85	.64**
4	32				27.14	7.56	.29
1	33				33.76	8.29	.12
2	34				36.78	7.84	.74**
2	**35**				35.44	8.43	.51*
4	36				36.04	7.96	.37
3	37				34.65	8.64	.82**
1	**38**				41.13	9.47	.13

	Test score	Self-efficacy	Correlation
Mean non-selected classes:	30.27	8.11	.43**
Mean selected classes:	34.59	8.29	.48**
Difference score:	4.32	0.19	
t-test:	3.88 (p<.001)	1.26 (n.s.)	

Legend:

Class selected for phase 2

Class did not partake

Based on my a priori selection criteria, where I included a) classes from regions 1 and 2, that participated in phase 1 in 2013, b) grade 10 classes that had participated during both years of phase 1, and c) classes with a range of mean test scores and self-efficacy scores during phase 1, I identified and invited eight classes to partake. The teacher of one grade-6 class declined (class number 28), thus an additional grade 6 class was invited and included (class number

21). Furthermore, the teacher of one grade-10 class (class number 38) informed us the last new task of the year would be introduced shortly after our first communication. As I was unsure whether I would be ready to collect data by then, an additional grade-10 class (class number 34) was also invited, however the teacher declined. Instead, a different class (class number 20) was invited, and included. In the end, I collected data from nine classes (highlighted yellow in Tables 29 and 30) in phase 2.

As can be seen in Table 29 the grade 6 classes selected for phase 2 had marginally stronger self-efficacy in grade 5 as compared with the classes that were not selected, while there were no significant differences in regard to mean national test scores. In regard to the older sample (Table 30), the classes selected for phase 2 had significantly higher scores on national tests, while there were no significant differences in regard to students' self-efficacy scores. Although I wished to include classes with a representative range of mean scores on national tests from phase 1, I also wished to include the three classes from regions 1 and 2 that participated in 2012 as well as 2013. Students in one of these classes (class 10) had relatively high mean scores on national tests during both 2012 and 2013. If for instance class 33 (the last grade 10 class that participated during 2013) had participated instead of class 10, the difference in mean national test scores between the selected and the non-selected classes would have been only marginally significant (difference score = 1.90, $p<.10$). In other words, I prioritised to focus on the possibility for studying individual learning trajectories over time, and thus ended up with a grade-10 sample with higher mean national test scores as compared with the total pool of phase 1 participants. This has implications for the interpretations of my phase 2 findings.

Some new students joined the selected classes after phase 1. Overall, 12 students participated in phase 2 who had not participated already in phase 1 of the study. In grade 6 there was one new student in every class, except for class number 31, where there were three new students. In grade 10 the new students were in class number 20 (four new students) and in class number 38 (two new students).

4.3.3 Procedure

I drew on experiences from the phase 1 data collection when planning studies 3 and 4. Prior to data collection, I sent the principal at each school a letter of information, and the principals generally discussed participation with the teachers directly. Once the teachers agreed, direct contact was established with myself via email. The teachers were encouraged to read the information letter carefully, and get in touch if they had any concerns or questions. Furthermore, I visited each school prior to data collection. During these visits I presented students with letters of participant information, and answered questions about the study. At most schools I was also present for one lesson in mathematics. Given the complex nature of the research design (see Figure 23), these visits were invaluable to minimize misunderstandings and plan logistics such as when the teacher would be presenting something new for the students, and at what times during the day the students could fill out the questionnaires. Also, this gave each teacher an opportunity to express their concerns or ask questions that they might not want to raise via an email or over the phone. About one week prior to data collection, I contacted the teachers again to hear about their plans for the coming lessons. At this stage the teacher shared the lesson topics, and the main task for each lesson, in order that the student questionnaires could be adapted for each class and lesson.

For each class, a timetable was worked out in collaboration with the teacher, thus the timings of pre-observation questionnaires and the time and number of the lessons I followed varied

from class to class. For example, some schools had "normal" 45-minute lessons (which is common in Norway), some had 60-minute lessons, and one grade 6 class had double lessons (2 x 45 minutes). For all classes, the students completed the pre-observation questionnaires prior to the first mathematics lessons followed. The researchers administered the pre-observation questionnaires, generally with the mathematics teacher present in the room. The questionnaires were first presented and explained on a projector or smart board. The students were asked to consider the "test" items (rating their confidence to lift bags of a certain weight) preceding the self-efficacy items, to become more familiar with the scoring format. The single self-efficacy task item from the end of lesson questionnaires was also demonstrated on the projector. The students were encouraged to ask for help from the researchers if they experienced difficulty with the questionnaires. In a couple of 6th grade classes the teacher and/or an assistant helped certain students to read the questionnaires – this had been arranged with the teacher beforehand. The teachers generally also answered questions from the students, but they were asked to only assist the students in understanding the items, and not tell them how to respond.

During the lessons in mathematics, the researchers generally sat at two separate desks towards the back of the classroom. Prior to the start of each lesson, they placed the questionnaires on the students' desks. In one of the first grade 6 classes visited the teacher "handed the lesson over" to the researchers when it was time for the students to fill out the questionnaire, so the principal researcher was the one to tell the students to fill out the first page of the questionnaires. In all later classes the teachers were explicitly encouraged by the researchers to take control of the start of the lesson themselves, so the researchers could take a more passive role, and focus on observing the lessons. The start of lesson questionnaires took about 3 minutes to complete, while the end of lesson questionnaires took about 5-10 minutes. The

questionnaires were in booklets, and stayed on the students' desks during the lessons. Students wrote their names on the front of the booklet for the purpose of data-linkage.

4.3.3.1 *Feedback for schools and teachers*

After data collection was completed, all the participating schools were offered to receive a presentation by the researchers about self-efficacy in general, why it is important in regard to classroom learning, and the theories behind some of the instruments used. All participating schools had requested and received such presentations by spring 2015. The nine participating teachers were also invited to ask for feedback from the researchers. It was emphasised we would not give feedback based on findings from this study, nor any information about specific students we had observed, and that any feedback would be related to theories about self-efficacy. Two teachers requested and were provided with such feedback. For this, we used our notes and observation data to provide examples of teacher behaviour that is likely to support adaptive self-efficacy (things the teacher was already doing well and could do more of), as well as teacher behaviour that may be less supportive of student self-efficacy (things the teacher could consider doing differently). We emphasised that our feedback was based on a review of previous studies of self-efficacy, and not related to our findings, as we had not yet completed analysing the data. The teachers were given the choice whether they preferred receiving this feedback in writing or in a phone conversation. Both teachers elected to receive written feedback.

4.3.4 Further ethical considerations

While I have thoroughly discussed my interactions with schools, teachers, and students in sections 4.3.3 and 4.3.3.1, I will here emphasise additional ethical considerations for the phase 2 data collection. First, given that the participants included children under the age of 18, letters of information were distributed to students and their guardians (Appendix L), and parental consent was collected. Information letters were also sent to the teachers (Appendix

M), via the principals. The letters of information provided details about the study, for instance that information would be treated confidentially, that participation was voluntary, and how I would handle any data.

As discussed in section 2.3.3, for any study the potential benefits from the findings should be considered relative to the inconvenience or potential harm to participants. In the case of studies 3 and 4, my data collection was of a relatively larger impact to schools / students as compared to that of phase 1. Each class was followed for 3-4 lessons; two observers were present during most lessons; children filled out questionnaires at the beginning and end of each of the lessons; and all teachers plus some children were interviewed several times. I took care to keep the questionnaires short, and emphasised teachers should not plan or do anything different during the lessons. All participating schools were offered and accepted visits by the researchers, with the aim to raise general awareness about self-efficacy with the teachers (see section 4.3.3.1). Overall, I argue the potential direct benefits to students, in regard to increased knowledge about how to support adaptive self-efficacy in mathematics, outweighed the inconvenience to the participants. Ethical approval was obtained from Norwegian authorities as well as the University of Oxford (Appendix F), and the information from phase 2 was handled with the similar care as the phase 1 data (see section 2.3.2).

4.3.5 Measures

I used data from two measures to answer the research questions in study 3. These include a classroom version and a short, classroom version of the SEGD (adapted from the SEGD, test version, developed for The National Test Phase of the DPhil, see section 2.5.1). The original measures (in Norwegian) are presented in Appendix R. I developed both of these measures myself. In addition, I collected data regarding students' ability conceptions (in the pre-observation questionnaire), their self-efficacy task expectations, and their mastery experiences

(self-efficacy task measure and mastery experiences measure both at the end of lessons 1-4), which will be used for the study 4 investigations in the present thesis. Further measures were used in order to investigate themes regarding classroom quality and students' self-efficacy, which will not be included in the present thesis. Since completing data collection I have decided to expand my focus on the methodological aspects of measuring self-efficacy, leading to a corresponding narrowing of themes, considering what I can reasonably pursue in-depth within the thesis. The additional measures included sources of self-efficacy and task experiences (in the end-of-lesson questionnaires), structured classroom observations, and teacher and student interview data. These "additional" measures will be used to investigate themes in future studies. I piloted the design of the phase 2 data collection three times, across a single lesson, including student questionnaires at the beginning and end of the lesson.

For the present study, the first measure was included in the student questionnaires administered prior to the first lesson observed.

Self-Efficacy Gradations of Difficulty (SEGD), classroom version (Appendix J)

This questionnaire was adapted from the SEGD, test version, developed for The National Test Phase (see study 1, section 2.5.1), including facets for problem solving and self-regulation in mathematics, and three levels of perceived difficulty. For phase 2, one additional facet was included: learning a new topic with a certain amount of help from the teacher, resulting in five facets of mathematics. The questionnaire was adapted to measure students' self-efficacy for the next sequence of lessons in mathematics. Students' expectations were targeted through an introductory statement at the beginning of the questionnaire (*The following sentences are related to the next few lessons in mathematics*), as well as introductory statements for each of the facets of mathematics (e.g. *During the next few lessons in mathematics you will learn a new topic. How certain are you that you can learn this new topic with a certain amount of*

help from the teacher?). These introductions were followed by statements of varying difficulty (e.g. ***During the next few lessons I can learn a new topic -*** *if I get a lot of help from the teacher*). For each item, students were asked to indicate their confidence on a 0-10 scale, with the anchors 0 = Not at all certain, 5 = Moderately certain, and 10 = Completely certain. Descriptives for this measure are presented in Table 31. It is important to note items 1-9 in the pre-observation questionnaires do not correspond with items 1-9 in the lesson questionnaires. Rather, items 1-3 are the same across both questionnaires, while items 7-9 in the pre-observation questionnaires correspond with items 4-6 in the lesson questionnaires, and items 13-15 in the pre-observation questionnaires correspond with items 7-9 in the lesson questionnaires.

Inspecting the descriptives in Table 31, we see students in both grades have higher mean self-efficacy for easy level tasks, followed by medium difficulty and hard tasks. The scores for grade 6 and grade 10 students are comparable, where the mean scores are higher among the grade 6 students for some items, and among the grade 10 students for other items. In the case of both groups, the easy items are more skewed and have higher kurtosis than the medium difficulty or hard items. This might be related to a so-called "ceiling effect" of the scores on the easy level items, as most students are able to complete the easiest tasks (see section 2.3.3). I used the robust maximum likelihood estimator in MPlus to deal with the non-normally distributed data. The Cronbach's alpha for the SEGD adapted version was .92.

Table 31: Descriptives Self-Efficacy pre-observations

Item	grade 6							grade 10						
	n	missing	M	SD	Range	Skew Z	Kurt Z	n	missing	M	SD	Range	Skew Z	Kurt Z
During the next few lessons I can learn a new topic…														
1 if I get lots of help from the teacher	74	7	8.39	2.38	10	-5.79	3.99	96	4	8.57	2.01	10	-6.68	6.39
2 if I get some help from the teacher	75	6	7.21	2.09	7	**-0.29**	-2.22	96	4	7.30	2.42	8	**-1.12**	-2.53
3 without any help from the teacher at all	75	6	3.56	2.71	9	**1.39**	**-1.60**	96	4	4.41	2.74	10	**0.13**	**-1.76**
During each of the next few lessons I can solve…														
4 at least some of the problems I attempt	75	6	8.84	2.19	10	-7.53	7.21	95	5	9.07	1.70	7	-7.48	4.90
5 at least half of the problems I attempt	75	6	7.89	2.18	9	-3.36	**0.25**	94	6	7.67	2.33	9	-2.46	**-1.34**
6 all of the problems I attempt	75	6	5.40	2.57	10	**-0.70**	**-0.92**	93	7	4.40	2.84	9	**-0.33**	-2.08
During the next few lessons I can solve…														
7 all the easy tasks	75	6	9.47	1.19	5	-9.81	13.00	95	5	9.24	1.29	6	-8.34	8.65
8 all the medium difficulty tasks	75	6	7.57	1.87	7	-2.09	**-0.69**	95	5	7.32	2.22	9	-2.28	**-0.68**
9 all the hard tasks	75	6	4.84	2.28	10	**-0.23**	**-0.77**	96	4	4.10	2.61	9	**-0.12**	-2.03
During the next few lessons I will be able to concentrate…														
10 at least a little while	75	6	9.27	1.61	7	-8.56	8.99	96	4	9.14	2.06	10	-12.11	18.24
11 at least half the time	75	6	8.37	2.05	8	-4.45	**1.42**	95	5	7.96	2.54	10	-5.00	1.97
12 the whole time	75	6	6.19	2.64	10	**-1.67**	**-1.14**	96	4	5.36	3.28	10	**-1.00**	-2.31
During the next few lessons I can manage *not to give up*…														
13 if I am faced with slightly difficult tasks	74	7	8.62	2.55	10	-8.15	8.45	96	4	8.39	2.57	10	-7.10	5.12
14 if I am faced with medium difficulty tasks	75	6	7.36	2.53	10	-4.32	2.28	96	4	7.54	2.64	10	-4.09	**0.75**
15 if I am faced with very difficult tasks	75	6	5.55	2.83	10	**-0.73**	**-1.62**	96	4	5.60	3.17	10	**-1.61**	**-1.77**
Valid N (listwise)	73							91						

Note: z-score is larger than 1.96 for all items, *except* those in **bold**

The student lesson questionnaires included a total of five measures, where students responded to one at the start and four at the end of each lesson. For the purposes of study 3, I use only the one measure from the beginning of the lesson.

Self-Efficacy Gradations of Difficulty (SEGD), short classroom version (Appendix K)

The short classroom version of the SEGD includes three of the five facets included in the full-length, adapted, version: problem solving, self-regulation and learning. The measure was adapted to each class through swapping one word referring to the specific topic (for instance geometry) that class pursued (e.g. *For each of these questions, think about the lesson on <geometry> which is just about to start*). The grade 10 topics were geometry (two classes) and equations (three classes). Three of the grade 6 topics were fractions, while one was measurement. The students responded to the SEGD short classroom version at the start of each lesson.

I piloted the short version of the SEGD three times with small groups of students in grade 6 to 8. The students filled out questionnaires while I was present, and were encouraged to ask questions. I conducted small group interviews asking the students about their interpretation of each item, if the items made sense to them, and their reasons for responding the way they did. I considered students' comments and also scanned their responses for inconsistencies, and any seemingly problematic items were adjusted accordingly. Any changes were then tested with a subsequent group of pilot students, over two cycles.

Descriptives for the SEGD for each lesson are presented in Tables 32 to 35. These indicate some of the same tendencies as observed in the pre-observation self-efficacy scores. For both student groups, mean scores are consistently higher for easy items, followed by medium

difficulty and hard items. Also, estimates of skewness and kurtosis are higher for the easy, as compared to the medium or hard items. When looking at Tables 32 to 35, more items have "problematic" estimates of skewness or kurtosis in the grade 6 than in the grade 10 sample. To account for the non-normally distributed data I used the robust maximum likelihood (MLR) estimator in Mplus (Lei & Wu, 2012). The Cronbach's alpha for the SEGD short version was .93, .92, .93, and .94, for each of the four lessons, respectively.

Table 32: Descriptives Self-Efficacy lesson 1

Item	grade 6							grade 10						
	n	missing	M	SD	Range	Skew Z	Kurt Z	n	missing	M	SD	Range	Skew Z	Kurt Z
In the next lesson in maths I can learn <geometry>…														
1 if I get lots of help from the teacher	77	4	8.97	1.92	10	-8.31	10.73	88	12	8.52	2.02	10	-6.12	5.50
2 if I get some help from the teacher	77	4	8.23	1.95	8	-4.53	2.42	88	12	7.49	2.48	9	-2.81	**-0.72**
3 without any help from the teacher at all	78	3	5.29	2.92	10	**-0.51**	**-1.66**	88	12	4.93	2.77	10	**-0.98**	**-1.94**
In the next lesson in maths I can solve…														
4 all the easy tasks	78	3	9.58	1.06	5	-12.50	22.18	87	13	8.95	1.66	9	-8.06	10.98
5 all the medium difficulty tasks	77	4	8.06	1.77	7	-2.71	**-0.20**	87	13	7.25	2.24	10	-2.36	**0.08**
6 all the hard tasks	77	4	5.52	2.38	10	**-0.85**	**-0.59**	88	12	4.55	2.56	9	**-0.81**	**-1.82**
In the next lesson in maths I can manage *not to give up*...														
7 if I am faced with slightly difficult tasks	77	4	8.82	2.13	10	-7.54	7.50	88	12	8.64	1.98	8	-6.24	3.78
8 if I am faced with medium difficulty tasks	77	4	8.05	2.31	10	-4.92	2.74	87	13	7.55	2.50	10	-2.95	**-0.44**
9 if I am faced with very difficult tasks	78	3	6.26	2.90	10	**-1.63**	**-1.23**	87	13	5.33	3.13	10	**-0.67**	-2.21
Valid N (listwise)	75							85						

Absolute value of skew / kurt higher than recommended cut-off

Note: z-score is larger than 1.96 for all items, *except* those in **bold**

Table 33: Descriptives Self-Efficacy lesson 2

Item	grade 6							grade 10						
	n	missing	M	SD	Range	Skew Z	Kurt Z	n	missing	M	SD	Range	Skew Z	Kurt Z
In the next lesson in maths I can learn <geometry>…														
1 if I get lots of help from the teacher	63	18	9.16	1.83	8	-7.94	8.41	87	13	8.94	1.94	10	-9.68	14.05
2 if I get some help from the teacher	63	18	8.87	1.57	6	-4.41	**1.56**	87	13	8.20	2.34	10	-5.27	2.93
3 without any help from the teacher at all	63	18	6.08	3.12	10	**-1.62**	**-1.39**	87	13	5.79	2.70	10	**-1.17**	**-1.17**
In the next lesson in maths I can solve…														
4 all the easy tasks	63	18	9.75	0.72	3	-10.24	15.15	87	13	9.25	1.67	10	-13.29	27.89
5 all the medium difficulty tasks	63	18	8.51	1.61	6	-4.29	**1.89**	87	13	7.91	2.38	10	-5.19	3.19
6 all the hard tasks	63	18	6.52	2.51	10	-2.45	**-0.07**	86	14	5.27	2.74	10	**-1.02**	**-1.65**
In the next lesson in maths I can manage *not to give up*...														
7 if I am faced with slightly difficult tasks	63	18	9.19	2.22	10	-10.34	15.71	87	13	8.95	2.30	10	-10.81	15.17
8 if I am faced with medium difficulty tasks	63	18	8.44	2.40	10	-5.81	4.36	87	13	8.05	2.58	10	-5.69	3.16
9 if I am faced with very difficult tasks	63	18	7.06	2.86	10	-2.38	**-1.20**	87	13	6.07	3.17	10	**-1.60**	-2.15
Valid N (listwise)	63							86						

Absolute value of skew / kurt higher than recommended cut-off

Note: z-score is larger than 1.96 for all items, *except* those in **bold**

Table 34: Descriptives Self-Efficacy lesson 3

Item	grade 6							grade 10						
	n	missing	M	SD	Range	Skew Z	Kurt Z	n	missing	M	SD	Range	Skew Z	Kurt Z
In the next lesson in maths I can learn <geometry>…														
1 if I get lots of help from the teacher	72	9	9.14	2.05	10	-9.87	13.73	87	13	9.00	2.05	10	-9.95	13.14
2 if I get some help from the teacher	72	9	8.79	1.92	10	-7.51	9.74	87	13	7.90	2.59	10	-4.78	**1.67**
3 without any help from the teacher at all	72	9	6.72	2.81	10	-2.23	**-1.17**	87	13	5.38	3.07	10	**-1.40**	-2.12
In the next lesson in maths I can solve…														
4 all the easy tasks	71	10	9.86	0.66	5	-21.94	77.10	87	13	8.94	1.94	8	-8.13	7.76
5 all the medium difficulty tasks	72	9	8.93	1.58	6	-5.53	2.72	86	14	7.63	2.53	9	-3.52	**-0.29**
6 all the hard tasks	71	10	6.82	2.53	10	-2.60	**-0.74**	87	13	5.55	2.80	10	**-1.79**	**-1.61**
In the next lesson in maths I can manage *not to give up*...														
7 if I am faced with slightly difficult tasks	72	9	9.36	1.83	10	-12.67	24.03	87	13	8.82	2.19	9	-7.60	5.97
8 if I am faced with medium difficulty tasks	72	9	8.82	2.07	10	-7.28	7.88	87	13	7.78	2.73	9	-4.37	**0.39**
9 if I am faced with very difficult tasks	72	9	7.47	2.65	10	-2.99	**-0.58**	87	13	6.16	3.37	10	**-1.79**	-2.12
Valid N (listwise)	71							86						

Absolute value of skew / kurt higher than recommended cut-off

Note: z-score is larger than 1.96 for all items, *except* those in **bold**

Table 35: Descriptives Self-Efficacy lesson 4

Item	grade 6							grade 10						
	n	missing	M	SD	Range	Skew Z	Kurt Z	n	missing	M	SD	Range	Skew Z	Kurt Z
In the next lesson in maths I can learn <geometry>…														
1 if I get lots of help from the teacher	57	24	9.30	1.72	7	-7.73	7.81	71	29	8.80	2.12	10	-7.27	7.78
2 if I get some help from the teacher	57	24	8.88	1.76	7	-5.35	3.38	71	29	7.83	2.66	10	-3.97	**0.76**
3 without any help from the teacher at all	57	24	6.74	2.77	10	-2.81	**-0.33**	71	29	5.55	3.08	10	**-1.35**	**-1.77**
In the next lesson in maths I can solve…														
4 all the easy tasks	57	24	9.91	0.34	2	-13.48	30.83	71	29	8.90	1.99	9	-7.16	6.81
5 all the medium difficulty tasks	57	24	9.16	1.22	5	-4.65	2.25	71	29	7.77	2.59	10	-3.47	**0.22**
6 all the hard tasks	57	24	7.07	2.38	9	-2.25	**-0.64**	71	29	5.49	2.87	10	**-1.34**	**-1.69**
In the next lesson in maths I can manage *not to give up*...														
7 if I am faced with slightly difficult tasks	56	25	9.75	0.96	5	-14.46	34.00	71	29	8.90	2.11	9	-6.95	5.48
8 if I am faced with medium difficulty tasks	55	26	9.31	1.41	7	-7.76	11.03	71	29	7.89	2.71	10	-3.59	**-0.23**
9 if I am faced with very difficult tasks	56	25	7.91	2.74	10	-3.99	**1.06**	71	29	6.24	3.29	10	**-1.63**	**-1.86**
Valid N (listwise)	55							71						

Absolute value of skew / kurt higher than recommended cut-off

Note: z-score is larger than 1.96 for all items, *except* those in **bold**

Table 36: Self-Efficacy mean, lesson by lesson

	Item	grade 6					grade 10				
		Preobs	L1	L2	L3	L4	Preobs	L1	L2	L3	L4
In the next lesson in maths I can learn <geometry>…											
1	if I get lots of help from the teacher	8.39	8.97	9.16	9.14	9.30	8.57	8.52	8.94	9.00	8.80
2	if I get some help from the teacher	7.21	8.23	8.87	8.79	8.88	7.30	7.49	8.20	7.90	7.83
3	without any help from the teacher at all	3.56	5.29	6.08	6.72	6.74	4.41	4.93	5.79	5.38	5.55
In the next lesson in maths I can solve…											
4	all the easy tasks	9.47	9.58	9.75	9.86	9.91	9.24	8.95	9.25	8.94	8.90
5	all the medium difficulty tasks	7.57	8.06	8.51	8.93	9.16	7.32	7.25	7.91	7.63	7.77
6	all the hard tasks	4.84	5.52	6.52	6.82	7.07	4.10	4.55	5.27	5.55	5.49
In the next lesson in maths I can manage *not to give up*...											
7	if I am faced with slightly difficult tasks	8.62	8.82	9.19	9.36	9.75	8.39	8.64	8.95	8.82	8.90
8	if I am faced with medium difficulty tasks	7.36	8.05	8.44	8.82	9.31	7.54	7.55	8.05	7.78	7.89
9	if I am faced with very difficult tasks	5.55	6.26	7.06	7.47	7.91	5.60	5.33	6.07	6.16	6.24
	Valid N (listwise)	75					85				

The absolute mean scores across all five time points are presented in Table 36. In the case of both grade 6 and grade 10, students' mean self-efficacy at the beginning of lesson four is higher than their pre-observation self-efficacy, in absolute terms. The exception is item 4; *I can solve all the easy tasks*, where grade 10 students' mean scores declined slightly, in absolute terms, across the measurement points. In general, the difference in scores between lesson four and the pre-observations was more pronounced in the case of grade 6, than for grade 10 students, indicating greater stability of self-efficacy for the older students. Finally, there is a tendency for students in both grades, that self-efficacy scores on the items related to the hard level tasks changed to a larger degree than the easy tasks. This can be seen when inspecting the mean absolute scores at lesson four, as compared to the pre-observation scores.

4.3.6 Analyses

The student questionnaire data was optically scanned and entered into Microsoft Excel electronically. I used an external agency for this (InfoShare Solutions AS). Any ambiguous student responses were flagged and visually inspected by me. Once the data was electronically entered, I used SPSS (version 24) for data management and to produce descriptive statistics, and Mplus (version 1.31 for mac: Muthén & Muthén, 2012) for factor and path analyses. Similar to the other studies in my thesis, I used pairwise deletion to handle missing data when producing correlation estimates in SPSS. Furthermore, I used the robust maximum likelihood (MLR) estimator in Mplus, which incorporates a full-information maximum likelihood approach to missingness (Graham & Coffman, 2012), for model estimations.

Similar to studies 1 and 2 in the present thesis, I used the following parameters and cut-offs to determine acceptable model fit: The chi square (χ^2/df <3 acceptable), the Root Mean Square Error of Approximation (RMSEA <.08 acceptable), the Comparative Fit Index (CFI) and the

Tucker-Lewis index (TLI) (CFI/TLI > .90 acceptable) the Standardized Root Mean Square Residual (SRMR<.10 acceptable), (Chiorri 2014; Morin, Marsh, Nagengast, and Scalas 2014; Schermelleh-Engel, Moosbrugger, and Müller 2003). For comparison between alternative models, I used cut-offs suggested by Chen (2007) to reject a more parsimonious model: ΔRMSEA (.015), ΔCFI (-.010), ΔSRMR (.030). In the case of nested models, and for tests of invariance, I considered both $\Delta\chi^2$ in relation to the difference in degrees of freedom (a non-significant $\Delta\chi^2$ indicating invariance/ support for the more parsimonious model) and ΔCFI (ΔCFI equal to or less than .010 indicating invariance/ support for the more parsimonious model). According to several researchers the $\Delta\chi^2$ can be too sensitive when considering invariance testing (e.g. Klassen et al., 2009; Little, 2013, p. 155), thus considering ΔCFI is recommended. Fit indices were thoroughly discussed in section 2.3.5.

Several of my latent growth models (research question 3) resulted in a negative residual variance on a single estimate. In this case I ran the series of tests recommended by Chen et al. (2001), in order to establish whether the negative residual was likely due to sample fluctuations. If the tests indicated it was reasonable, I restricted the residual variance of the "offending" item above 0. Negative residuals, or "Heywood cases" as they are commonly referred to as, were discussed in section 2.3.5, study 1. In the next section I will describe how I calculated "reliable change" scores, before I present my analytic strategy for my research questions.

4.3.6.1 Calculating reliable change

In order to assess variability in students' self-efficacy scores as an effect of levels of difficulty, I first calculated the "differential continuity" of the self-efficacy scores. This includes the correlation estimates between self-efficacy scores at different time points, and is the same kind of test used for test-retest analyses. Accordingly, a threshold of r=.85 is used to

indicate whether a construct is reliable across observations. Bernacki et al. (2015) argued that estimates below r=.85 indicated there was variability in the self-efficacy scores beyond what might be expected due to measurement error. I investigated the differential continuity for three self-efficacy level constructs (self-efficacy for easy, medium, and hard level tasks) measured at the beginning of each lesson.

In order to investigate intra-individual change in self-efficacy in relation to levels of difficulty, I calculated the "Reliable Change" score (Bernacki et al., 2015; Christensen & Mendoza, 1986; Jacobson & Truax, 1991). This is the difference in self-efficacy scores across two time points, contrasted with the standard error of the difference score. The formulae is $RC= (X^2-X^1) / S_{diff.}$, where RC= reliable change, X^1= self-efficacy score at time 1, X^2= self-efficacy score at time 2, and $S_{diff.}$= standard error of difference between the two self-efficacy scores (Christensen & Mendoza, 1986). When reliable change scores are larger than +/- 1.96, this indicates the obtained difference score would occur due to measurement error less than 5% of the time, in other words the score is likely to reflect "real change" in the construct rather than measurement error. Reliable change is an estimate traditionally used in psychology. The premise for the estimation is to investigate to what degree changes in scores (due to, for instance, an intervention or a treatment: in my case, changes in relation to lesson experiences) reflect real change, beyond what might be "statistically significant", or even of a certain effect size. Bernacki et al. (2015) used this score to estimate what percentage of students had experienced a reliable increase, decrease, or no change in self-efficacy from one measurement point to another. While Bernacki used a single repeated self-efficacy item, I used a multidimensional self-efficacy measure, and calculated the reliable change scores for each of the three levels of difficulty included in my measure. I included these calculations to

allow for comparability with the Bernacki et al. (2015) study, and in order to investigate differences in reliable change for varying levels of task difficulty.

Reliable change scores provide information about how many students reported self-efficacy changes of a certain "size", however it is a simple descriptive statistic (with a set cut-off point), it does not take into account measurement error of my proposed model, and it is not well suited to account for trends in the data. For instance, it is possible many students reported scores very close to the threshold, but were not included in the count. As such, I used structural equation modelling in MPlus to investigate these issues statistically. I next describe my analytic strategy to investigate the structural validity of my measure, and the stability and change of self-efficacy over time.

4.3.6.2 Analytic strategy

In order to investigate the structural validity of my measure, and the stability and change of students' self-efficacy across several time points, I developed a modelling strategy. In short, this involved establishing the factor structure of my self-efficacy measure (research question 1), testing for configural, weak and strong invariance over time (e.g. whether the construct I measured changed structure or meaning over time), and finally applying the factor structure to investigate my substantive research questions, using autoregressive path models (research question 2) and latent growth curve models (research question 3).

I chose to include grade 6 and grade 10 in the same model specifications for research questions 1 and 2. There are several reasons for this. First, my previous investigation into the factor structure of the SEGD indicated the factor structure remained invariant across the two samples, although some of the estimates differed (see study 1, section 2.5.1). As such, I presume grade might have less bearing on the factor structure, while it may influence the

substantive relationship between constructs, e.g. stability or change over time. In research question 2 I chose to include grade as a covariate, in order to alleviate for this. Second, my sample for phase 2 is substantially smaller than for phase 1, as the focus is on following students over time. Accordingly, several of the models did not converge when specified with a smaller sample (e.g. about half the current sample size). Third, there is an advantage to including grade as a covariate, as I chose to do for research question 2, as it enables statistical investigation of the influence by grade on students' self-efficacy. It is possible however there are grade-level differences (as well as individual differences) to the stability of students' self-efficacy, and thus I must take appropriate caution when interpreting my results. In the case of research question 3 I investigated the grade levels in separate models, which converged. This enabled specification of different growth shapes for the two samples, while comparing the effect of grade on the parameters was less straightforward (see section 4.4.4.2 for my approach to this). I will now go into detail about each of my modelling steps.

Analytic strategy research question 1

As a first step to investigating the structural validity of my measure (research question 1) I inspected the correlation matrices for the SEGD measure from the pre-observations and each of the lesson time point. I then specified a series of alternative measurement models. I tested the *a priori* hypothesized model for my self-efficacy measure (Model 2, see Figure 26), building on a review of the literature as well as the findings from studies 1 and 2 in this thesis (see sections 2.5.1 and 3.5.1). I specified also plausible alternative models, as well as a unidimensional model, for comparison purposes (see Figures 24 to 29).

Three of the proposed models were versions of the measurement models included in study 1 of the present thesis: the Unidimensional Model (Model 0), where item scores were predicted

by a single latent self-efficacy construct; the Latent Facets Model (Model 1), including three latent facet constructs and correlated uniquenesses for levels of difficulty; and the Latent Levels Model (Model 2), my *a priori* hypothesized measurement model, including three latent level constructs and correlated uniquenesses for facets. Three additional models were tested, that were not included in study 1. These were all latent levels Bi-factor Models (Models 3-5), proposed as alternatives to my hypothesised model (Model 2). The three bi-factor models represent alternative ways of dis-attenuating the construct of self-efficacy from varying perceptions of difficulty: The "Easy Self-Efficacy" construct (Model 3) represents students' self-efficacy for easy level tasks, dis-attenuated for their perceptions of harder levels of task difficulty, while the "Medium difficulty" and "Hard difficulty" constructs in Model 3 represent students' perceptions of medium and hard level tasks, respectively, dis-attenuated for their self-efficacy for easy level tasks. The same is true for the "Medium Self-Efficacy" and the "Hard Self-Efficacy" Bi-factor Models (Models 4 and 5). Except for the baseline, Unidimensional Model (Model 0), all my empirical models accounted for the proposed multidimensionality of the SEGD measure through including correlated uniquenesses (see study 1, section 2.3.4).

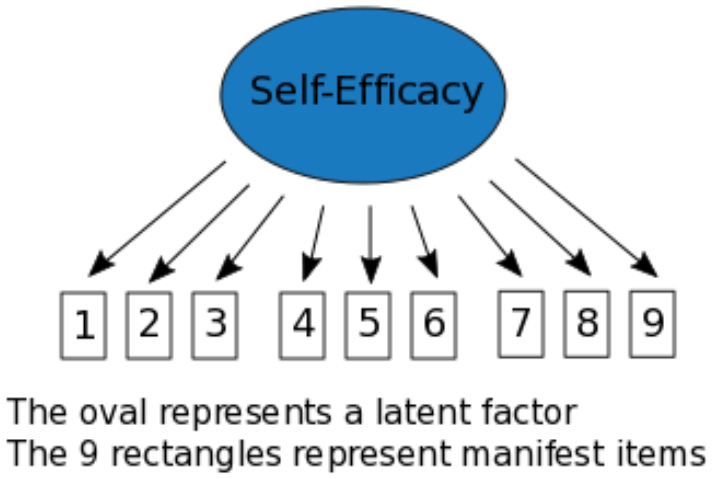
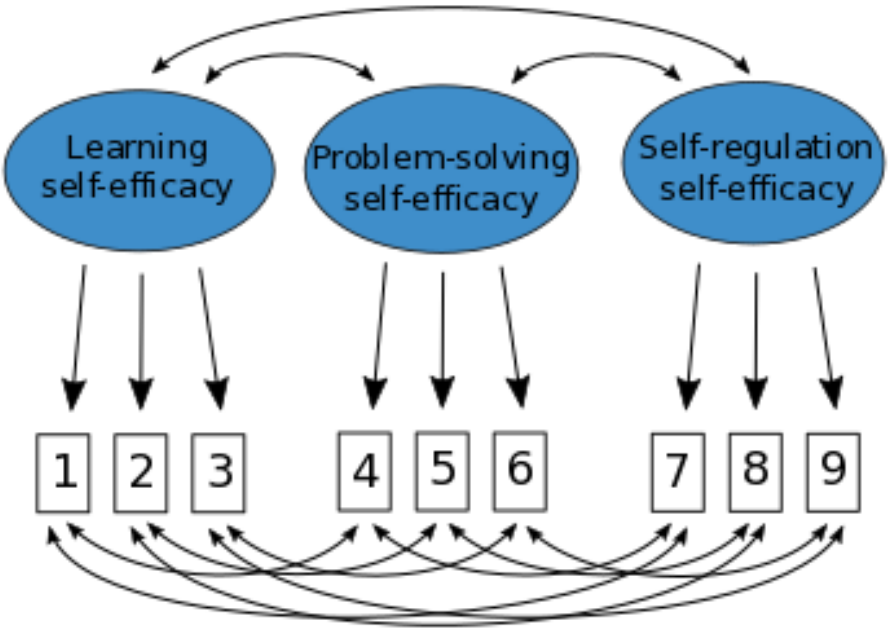
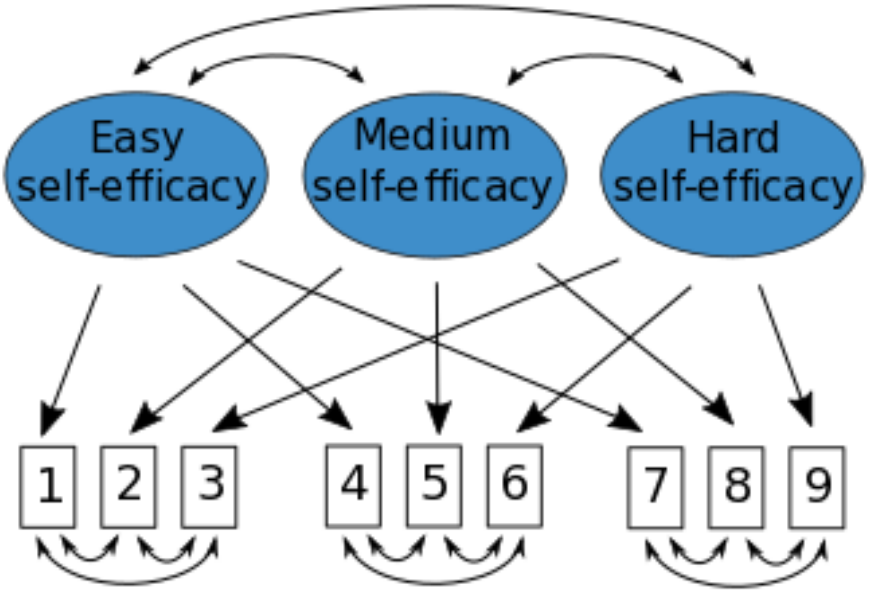


Figure 24: Model 0, Unidimensional



The 3 ovals represent latent facet factors
The 9 rectangles represent manifest items
Double-headed arrows between manifest items represent correlated uniquenesses (CU's) for levels of perceived difficulty

Figure 25: Model 1, Latent Facets



The 3 ovals represent latent factors
The 9 rectangles represent manifest items
Doubleheaded arrows between manifest items represent correlated uniquenesses (CU's) for facets of mathematics

Figure 26: Model 2, Latent Levels

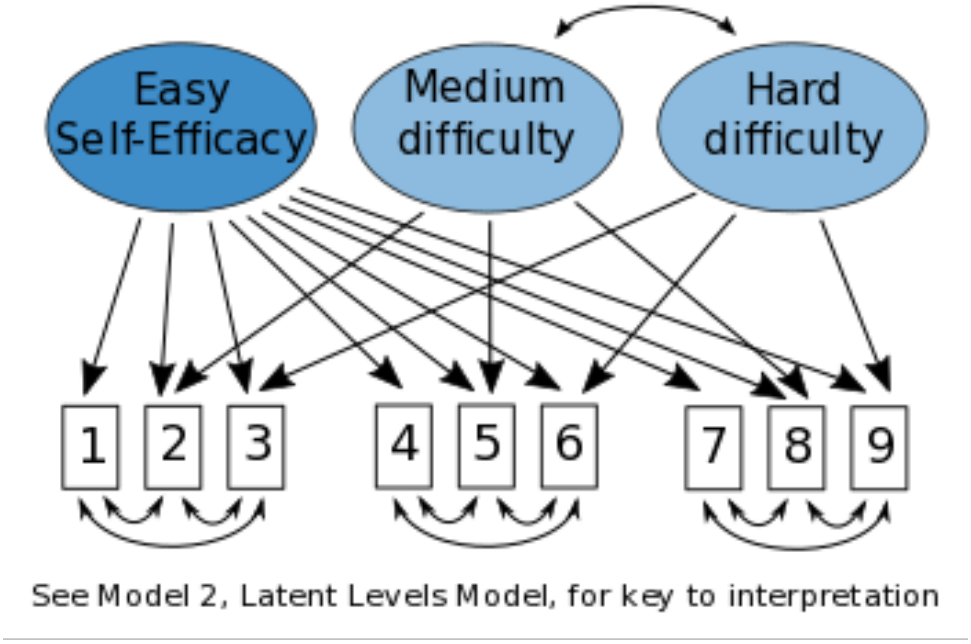


Figure 27: Model 3, Bi-factor Easy

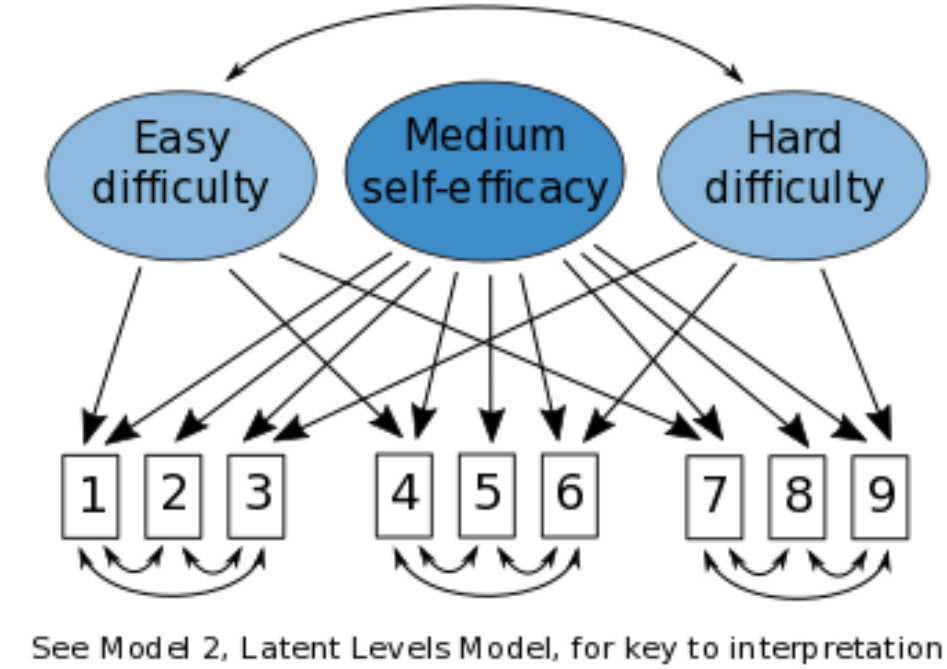


Figure 28: Model 4, Bi-factor Medium difficulty

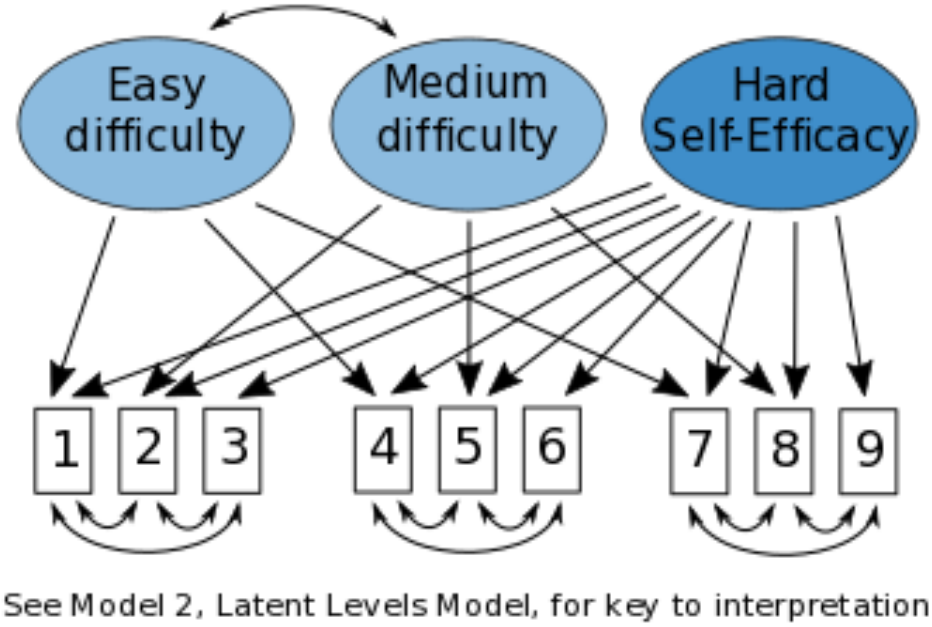


Figure 29: Model 5, Bi-factor Hard

My next concern was to establish configural invariance of my measure, that is, whether the same factor structure (measurement model) applied to the measure of self-efficacy at each time point. I investigated this through specifying for each time point each of the alternative empirical models using confirmatory factor analysis, to see whether my theoretical (hypothesised) model held at each time point (e.g. whether it lead to acceptable fit indices, according to the chosen cut-offs, as well as meaningful parameter estimates, and low standard errors). I presented the parameters and cut-offs I chose to assess model fit in section 4.3.6.

Having established the best fitting factor structure across each time point and tested for configural invariance (research question 1), I then proceeded to test for weak and strong invariance of my measure. Testing for weak invariance involves restricting the individual factor loadings to be equal across measurement points. Weak invariance is necessary for investigating regression parameters. Testing for strong invariance involves including also (in addition to weak invariance constraints) equality constraints for the factor intercepts across

the measurement points. If the models hold, this supports the notion that what is being measured is the same at each time. I compared models with increasing equality constraints to test whether the factor loadings and factor intercepts were invariant across time. Strong invariance is necessary for investigating mean-parameters. Given the complexity of my measurement model, I was unable to test for factorial invariance in a model where all my three latent level constructs were included. As an alternative, I specified separate models for the easy, medium, and hard latent self-efficacy constructs.

Analytic strategy research question 2

I next used the preferred measurement model (research question 1) to investigate the stability of the rank-order of students' self-efficacy across time (research question 2). I did this through testing for each of the three proposed self-efficacy constructs a basic simplex autoregressive model, before including this model in a path model including also grade and students' mathematics self-efficacy (as measured prior to the lessons) as covariates. The simplex model reflects the expected change process when change in a construct is steady and gradual, and the influences of any previous time points are fully mediated through the immediately preceding time point (Little, 2013, p. 184). Autoregressive models provide information about the stability of the rank-order of individuals across time: in the case of this study, whether students retain their rank-order (based on their self-efficacy) across time points. An advantage of autoregressive models is that they can easily accommodate several processes, and be used to investigate the relationship between them.

Analytic strategy research question 3

In order to investigate self-efficacy change across time (research question 3), I used latent growth curve modelling. Several alternative approaches exist for modelling change over time

in a structural equation model framework, for instance autoregressive models (which I used in research question 2), multilevel models for change (Little, 2013; Mehta & Neale, 2005), and continuous time modelling (Driver & Voelkle, 2016). The approaches differ in regard to what parameters are estimated, and as such, what kinds of substantive aims they can serve. Latent growth models and multilevel models for change can both be referred to as "random-coefficient multi-level models for nested data" (Little, 2013, p. 246), and provide information about the intercept (e.g. mean score at the starting point) and slope (e.g. growth trajectory) of a measure across time, as well as whether there are (significant) individual differences to these parameters. Given my focus on mean level changes across time, either of these two approaches might be reasonable. While both latent growth curve models and multilevel models for change are relatively well known and established, continuous time modelling is a novel approach developed within the R framework. Continuous time modelling is an interesting new approach, however it is still in its early days, and is for now restricted to a specific software; R. Given that my data collection was not designed with this analytic procedure in mind, it is beyond the scope of the present thesis to use this modelling approach. I chose latent growth models as they provide the possibility to compare different shapes (slopes) of growth in students' self-efficacy, which is of substantive interest.

Latent growth curve models are based on the assumption that distances between the measurement occasions are equal across individuals. I do not have equal distances between measurement points in my data. While most of the lessons were up to 24 hours apart, the time lag between the beginnings of two adjacent mathematics lessons varied from 45 minutes to 5 days (in the case of one class only). The effect of violating this assumption could cause bias in parameter estimates, particularly in the shape of the growth parameter. This has implications for my analyses, in particular I must consider these non-equidistant time lags when comparing

model fits of models with different growth shapes. This issue is also related to the degree to which self-efficacy change is related to the passing of time (e.g. across long time-lags, through student cognitive development), or to the experiences of events. In a theoretical sense, it is plausible changes in self-efficacy differ as a function of the distance between the measurement occasions, or rather, as a function of memory (how salient the last lesson in mathematics is). However, my review of the literature indicated self-efficacy changes were mainly related to events or experiences (interpreted by individuals), rather than mainly a linear function of time/cognitive development. Regardless, the violation of equidistant time points is a limitation of the present study, and has important implications in regard to interpretation of the findings, in particular in regard to the slope parameters. Specifically, the change (slope) parameters from research question 3 in the present study must be interpreted in terms of change across a number of sequential lessons, rather than change across a given time period.

Given the complexity of my hypothesised measurement model (Model 2 in this section: see also study 1, in The National Test Phase), it was likely a latent growth model based on item scores for several time points might either not converge, or result in poor fit. Little (2013, p. 263) describes how fit indices can decrease when using latent growth models as compared with confirmatory factor analyses. Indeed, a latent growth model for Model 2 (based on item scores, including 4 time points) did not result in 'acceptable' fit indices, and in addition, resulted in several negative residual variances. In order to check whether these "unacceptable" results were due to my chosen model or due to the complexity of the repeated measures data, I specified as latent growth models also a unidimensional model (Model 0), and the best-fitting alternative (plausible) models (Models 3-5). Neither of these models resulted in acceptable fit indices. This indicated the problem was not the measurement model

per se, but rather the complexity of testing a complex model structure across several time points. I thus decided to use the factor structure derived from the best fitting measurement model to create item parcels for each of the latent level constructs, in order to reduce the complexity of the models and the number of parameters estimated.

Parcelling as a modelling technique has been critiqued as it might hide (confound) the true factor structure of a measure (Bandalos, 2002). As such it is not recommended when establishing the factor structure of a new measure is the aim. For research question 3 of the present study, the aim was to investigate the relationship between latent constructs (e.g. latent means over time), and the factor structure of my self-efficacy measure had been investigated in research question 1. As such, parcelling technique was seen as a useful strategy to investigate the mean changes in a latent construct over time (e.g. Little et al., 2002). In line with my analytic strategy for research question 2 (stability of self-efficacy using autoregressive models), I specified separate latent growth models for each of the three proposed self-efficacy level constructs, to investigate changes in mean parameters over time.

4.4 Results

In this section, I first present descriptive statistics investigating reliable change of my self-efficacy constructs, followed by results of the factor analyses, investigating the structural validity of my self-efficacy measure (research question 1). I then present the results from my investigation into the stability of self-efficacy, using autoregressive models (research question 2), followed by my investigation into mean-level changes of self-efficacy over time, using latent growth models (research question 3).

4.4.1 Reliable change statistics

In order to investigate whether the variability of self-efficacy differed according to level-specificity, I investigated differential continuity, before I calculated the reliable change scores for the three proposed self-efficacy level constructs. The self-efficacy level constructs were based on the hypothesised measurement model (Model 2), which again is based on the preferred factor structure from study 1 in the thesis. The structural validity of this measurement model is investigated in research question 1, section 4.4.2, in the present study.

4.4.1.1 Differential continuity

In order to investigate differential continuity (e.g. stability, or indications of group level changes in self-efficacy over time) for each of the level constructs, I inspected the correlations between the proposed latent self-efficacy level constructs across the five time points. These estimates are presented in Table 37. Only few autocorrelations are above $r = .85$ in the case of grade 6 (two estimates), while substantially more are above $r = .85$ in the case of grade 10 (17 estimates). This indicates that the self-efficacy of the older students in my sample was more stable as compared to the younger students. As a matter of fact, the average correlation estimate was .55 for the grade 6 sample and .73 for the grade 10 sample, a difference which was statistically significant at $p<.001$ (investigated through an independent samples t-test). Overall, the correlations indicate a degree of variability in the self-efficacy scores across time points, as most autocorrelations are below the threshold of $r=.85$. This warrants a further investigation into intra-individual change in self-efficacy, which I first explored through the descriptive statistic "reliable change", presented next.

Table 37: Correlation matrix Self-Efficacy level constructs, pre-obs. - lesson 4

	Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	Pre-obs. Self-Efficacy easy tasks	1	.86**	.67**	.73**	.74**	.63**	.75**	.77**	.56**	.77**	.78**	.62**	.77**	.72**	.51**
2	Pre-obs. Self-Efficacy medium diff. tasks	.74**	1	.86**	.69**	.87**	.81**	.63**	.82**	.70**	.66**	.82**	.74**	.68**	.79**	.65**
3	Pre-obs. Self-Efficacy hard tasks	.43**	.82**	1	.50**	.75**	.81**	.45**	.71**	.74**	.45**	.68**	.76**	.45**	.62**	.63**
4	L1 Self-Efficacy easy tasks	.66**	.51**	.32**	1	.88**	.67**	.69**	.72**	.54**	.80**	.80**	.61**	.83**	.82**	.62**
5	L1 Self-Efficacy medium difficulty tasks	.54**	.70**	.60**	.76**	1	.88**	.66**	.81**	.70**	.69**	.83**	.72**	.75**	.85**	.74**
6	L1 Self-Efficacy hard tasks	.34**	.64**	.75**	.53**	.78**	1	.51**	.73**	.82**	.52**	.72**	.78**	.54**	.71**	.80**
7	L2 Self-Efficacy easy tasks	.55**	.46**	.26*	.74**	.65**	.43**	1	.89**	.59**	.91**	.81**	.55**	.91**	.83**	.52**
8	L2 Self-Efficacy medium difficulty tasks	.48**	.61**	.53**	.62**	.87**	.70**	.72**	1	.82**	.84**	.92**	.77**	.82**	.91**	.72**
9	L2 Self-Efficacy hard tasks	.22	.48**	.58**	.38**	.69**	.82**	.49**	.85**	1	.55**	.77**	.86**	.49**	.71**	.83**
10	L3 Self-Efficacy easy tasks	.44**	.38**	.22	.66**	.56**	.40**	.70**	.60**	.42**	1	.88**	.64**	.92**	.82**	.55**
11	L3 Self-Efficacy medium difficulty tasks	.59**	.57**	.38**	.71**	.74**	.57**	.60**	.80**	.63**	.75**	1	.87**	.84**	.93**	.78**
12	L3 Self-Efficacy hard tasks	.33**	.48**	.50**	.46**	.59**	.75**	.37**	.70**	.84**	.41**	.71**	1	.59**	.78**	.85**
13	L4 Self-Efficacy easy tasks	.33*	.41**	.36*	.37**	.36**	.36**	.74**	.63**	.56**	.60**	.54**	.33*	1	.90**	.63**
14	L4 Self-Efficacy medium difficulty tasks	.15	.59**	.66**	.21	.64**	.63**	.42**	.78**	.74**	.45**	.60**	.57**	.63**	1	.85**
15	L4 Self-Efficacy hard tasks	.19	.57**	.71**	.19	.48**	.77**	.22	.67**	.80**	.25	.57**	.86**	.48**	.76**	1

**Correlation is significant at the 0.01 level (2-tailed), *Correlation is significant at the 0.05 level (2-tailed).

Pre-obs. = Pre-observations, L1- L4 = Lesson 1 - Lesson 4

Grade 6 estimates are displayed below the diagonal, grade 10 above

4.4.1.2 Reliable change

In Table 38 I present the reliable change scores for the composite measures of easy, medium, and hard levels of self-efficacy, between each time point. It is worth noting the percentages are based on very few students, so these are simple descriptive scores and can not be taken as indications of general trends in the wider population. For grade 6, less than 20% of the students reported a reliable change in their self-efficacy between any time point, while for grade 10 this number was less than 10%. While the numbers of reliable changes were pretty similar across lessons and levels of difficulty in the case of grade 10, grade 6 students reported more reliable changes to their self-efficacy between the "pre-observation" time point and the first lesson, than between subsequent lessons. These changes were predominantly positive. Furthermore, more grade 6 students experienced changes in self-efficacy for hard, versus easy and medium level tasks. It is worth noting that the estimates for the self-efficacy level constructs are based on aggregated measures, thus less variation might be expected as compared to single item scores.

Table 38: Reliable increase, decrease or no change in Self-Efficacy

Self-Efficacy constructs		Grade 6								Grade 10							
		Pre-obs.- L1		Lesson 1-2		Lesson 2-3		Lesson 3-4		Pre-obs.- L1		Lesson 1-2		Lesson 2-3		Lesson 3-4	
		Freq.	%	Freq.	%	Freq.	%	Freq.	%	Freq.	%	Freq.	%	Freq.	%	Freq.	%
Easy level	Reliable increase	4	5.63%	2	3.28%	3	5.36%	2	3.92%	2	2.44%	1	1.22%	4	4.94%	4	6.06%
	Reliable decrease	1	1.41%	2	3.28%	2	3.57%	1	1.96%	4	4.88%	3	3.66%	2	2.47%	1	1.52%
	No change	66	92.96%	57	93.44%	51	91.07%	48	94.12%	76	92.68%	78	95.12%	75	92.59%	61	92.42%
Medium level	Reliable increase	7	9.72%	4	6.67%	4	7.02%	4	7.02%	1	1.25%	3	3.70%	2	2.50%	5	7.69%
	Reliable decrease	1	1.39%	0	0.00%	1	1.75%	1	1.75%	6	7.50%	1	1.23%	4	5.00%	0	0.00%
	No change	64	88.89%	56	93.33%	52	91.23%	52	91.23%	73	91.25%	77	95.06%	74	92.50%	60	92.31%
Hard level	Reliable increase	11	14.86%	6	9.84%	4	7.14%	3	5.77%	1	1.23%	4	4.88%	4	5.00%	1	1.52%
	Reliable decrease	3	4.05%	2	3.28%	0	0.00%	3	5.77%	2	2.47%	3	3.66%	3	3.75%	2	3.03%
	No change	60	81.08%	53	86.89%	52	92.86%	46	88.46%	78	96.30%	75	91.46%	73	91.25%	63	95.45%

Table 39 displays the number of students who reported one or more reliable changes to their self-efficacy across the time points, for easy, medium, and hard levels of difficulty respectively. Again, these descriptives indicate that more grade 6 students than grade 10 students experienced reliable changes to their self-efficacy. The difference is especially prominent for students' self-efficacy for hard level tasks, where nearly 30% of grade 6 students reported at least one reliable change across the five time points (pre-observations to lesson 4), compared with 15% of grade 10 students.

Table 39: Reliable change in Self-Efficacy across all time points

Self-Efficacy constructs		Grade 6		Grade 10	
		Freq.	%	Freq.	%
Easy level	No reliable change	70	86.42%	86	86.00%
	1 reliable change	7	8.64%	8	8.00%
	2 reliable changes	2	2.47%	5	5.00%
	3 reliable changes	2	2.47%	1	1.00%
Medium level	No reliable change	64	79.01%	81	81.00%
	1 reliable change	14	17.28%	16	16.00%
	2 reliable changes	2	2.47%	3	3.00%
	3 reliable changes	1	1.23%	0	0.00%
Hard level	No reliable change	57	70.37%	85	85.00%
	1 reliable change	19	23.46%	11	11.00%
	2 reliable changes	2	2.47%	3	3.00%
	3 reliable changes	3	3.70%	1	1.00%

4.4.1.3 Summary of reliable change descriptives

Reliable change scores were investigated for three proposed latent self-efficacy constructs related to easy, medium, and hard difficulty tasks. Differential continuity (e.g. autocorrelations among latent constructs across time) for students' easy, medium, and hard self-efficacy indicated there was sufficient change in the constructs across time to warrant further investigations. The autocorrelations were higher in the case of the older students, indicating higher stability across time as compared with the younger students.

Reliable change scores indeed indicated that for both grades, there were fluctuations in self-efficacy. Some students reported reliable decreases, while others reported reliable increases. The reliable change scores differed (in absolute terms) for self-efficacy associated with different levels of difficulty, as well as between the samples. In the case of grade 6, a higher percentage of students reported no reliable changes for the easy (86%), than for the medium (79%), than for the hard tasks (70%), indicating higher stability in the easy level self-efficacy construct. For the grade 10 sample, a similar percentage of students reported no reliable changes in relation to easy (86%) and hard level (85%) tasks, while a slightly lower percentage reported no reliable changes in regard to the medium difficulty (81%) tasks. This indicates somewhat higher stability in the medium level self-efficacy construct in the case of the grade 10 students. Looking across lessons, for grade 6, there were more reliable changes between the earlier time points, than the later ones. No such pattern was evident in the case of the grade 10 students.

4.4.2 RQ1: The structural validity of my lesson self-efficacy measure

4.4.2.1 Correlation matrices

I first inspected the item-level correlation matrices as a step to investigate the factor structure of the (multidimensional) adapted Self-Efficacy Gradations of Difficulty (SEGD) measure. I focused on the correlation matrices at each individual time point (a total of five correlation matrices), and I inspected correlation matrices for grade 6 and grade 10 scores separately. The correlation matrices for the adapted SEGD measure for each lesson, as well as the pre-observations, are presented in Tables 40 to 44. For purposes of readability, the term "I can manage not to give up" is swapped for the term "I can persevere" in the following tables. While the Norwegian questionnaire directly translated read "manage not to give up", the English translation is closer to the term persevere.

Some tendencies are evident at each of the five time points: the intercorrelations indicate students' self-efficacy is related to both facet specificity (indicated by cross-correlations within facets), and levels of perceived difficulty (indicated by higher correlations between items of closer levels of difficulty, e.g. easy and medium, or medium and hard, and smaller between easy and hard). All non-significant cross-correlations within each time point include items in different facets, and/or different levels of difficulty (indicating support for a multidimensional measure, in line with Study 1, section 2.5.1). The item with fewest significant cross-correlations is item 1, (*I can learn a new topic if I get lots of help from the teacher*). In grade 6, 15 correlations (of 46 possible) for this item were non-significant, looking within all time points. In grade 10 the same item was associated with five non-significant cross-correlations, all at the first time point (pre-observations). In line with the findings in section 4.4.1.1, the correlation estimates in grade 10 are higher than in grade 6, and there are more non-significant cross-correlations in grade 6 than in grade 10.

Table 40: Correlation matrix Self-efficacy, pre-observations

	Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	I can learn a new topic if I get a lot of help from the teacher	1	.49**	.39**	.20	.42**	.30**	.28**	.22*	.12	.18	.30**	.21*	.20	.22*	.17
2	I can learn a new topic if I get some help from the teacher	.41**	1	.58**	.54**	.61**	.42**	.53**	.62**	.52**	.39**	.47**	.29**	.42**	.53**	.48**
3	I can learn a new topic without any help from the teacher at all	.34**	.52**	1	.47**	.65**	.63**	.45**	.67**	.71**	.36**	.46**	.34**	.45**	.51**	.55**
4	I can solve at least some of the problems I attempt	.46**	.35**	.30**	1	.64**	.42**	.55**	.61**	.51**	.46**	.56**	.44**	.62**	.58**	.52**
5	I can solve at least half the problems I attempt	.24*	.44**	.31**	.72**	1	.78**	.58**	.79**	.74**	.38**	.61**	.53**	.64**	.75**	.66**
6	I can solve all the problems I attempt	.01	.37**	.28*	.12	.60**	1	.47**	.76**	.83**	.31**	.55**	.57**	.50**	.66**	.66**
7	I can solve all the easy tasks	.21	.36**	.00	.54**	.59**	.36**	1	.73**	.51**	.55**	.48**	.40**	.56**	.57**	.49**
8	I can solve all the medium difficulty tasks	.13	.41**	.26*	.46**	.73**	.66**	.64**	1	.84**	.50**	.64**	.50**	.66**	.77**	.71**
9	I can solve all the hard tasks	.16	.42**	.35**	.24*	.56**	.70**	.42**	.84**	1	.38**	.55**	.52**	.53**	.69**	.72**
10	I will be able to concentrate at least a little while	.22	.36**	.14	.46**	.33**	.08	.51**	.43**	.29*	1	.74**	.48**	.49**	.47**	.35**
11	I will be able to concentrate at least half the time	.11	.43**	.19	.35**	.40**	.31**	.49**	.49**	.38**	.79**	1	.80**	.53**	.59**	.48**
12	I will be able to concentrate all the time	.18	.37**	.15	.33**	.42**	.30**	.32**	.35**	.33**	.47**	.77**	1	.33**	.46**	.48**
13	I can <persevere> if I am faced with slightly difficult tasks	.13	.42**	.32**	.23*	.12	-.03	.18	.07	.03	.38**	.32**	.17	1	.87**	.55**
14	I can <persevere> if I am faced with medium difficulty tasks	.19	.45**	.39**	.32**	.25*	.12	.17	.20	.17	.33**	.38**	.32**	.87**	1	.75**
15	I can <persevere> if I am faced with very difficult tasks	.12	.36**	.46**	.32**	.37**	.30**	.16	.36**	.39**	.15	.25*	.29*	.47**	.79**	1

**Correlation is significant at the 0.01 level (2-tailed), *correlation is significant at the 0.05 level (2-tailed).

Grade 6 estimates are displayed below the diagonal, grade 10 above

Table 41: Correlation matrix Self-efficacy, lesson 1

	Item	1	2	3	4	5	6	7	8	9
1	I can learn <topic> if I get a lot of help from the teacher	1	.83**	.50**	.63**	.57**	.42**	.65**	.63**	.47**
2	I can learn <topic> if I get some help from the teacher	.30**	1	.74**	.75**	.82**	.67**	.73**	.80**	.69**
3	I can learn <topic> without any help from the teacher at all	.43**	.58**	1	.63**	.81**	.83**	.63**	.79**	.76**
4	I can solve all the easy tasks	.23*	.50**	.26*	1	.79**	.57**	.74**	.74**	.58**
5	I can solve all the medium difficulty tasks	.45**	.56**	.56**	.55**	1	.87**	.73**	.86**	.80**
6	I can solve all the hard tasks	.29*	.36**	.64**	.26*	.78**	1	.56**	.75**	.84**
7	I can persevere if I am faced with slightly difficult tasks	.41**	.62**	.37**	.34**	.42**	.32**	1	.86**	.65**
8	I can persevere if I am faced with medium difficulty tasks	.32**	.68**	.47**	.40**	.60**	.50**	.84**	1	.84**
9	I can persevere if I am faced with very difficult tasks	.27*	.49**	.52**	.33**	.62**	.63**	.51**	.80**	1

**Correlation is significant at the 0.01 level (2-tailed), *correlation is significant at the 0.05 level (2-tailed).

Grade 6 estimates are displayed below the diagonal, grade 10 above

Table 42: Correlation matrix Self-efficacy, lesson 2

	Item	1	2	3	4	5	6	7	8	9
1	I can learn <topic> if I get a lot of help from the teacher	1	.87**	.36**	.72**	.63**	.41**	.69**	.61**	.44**
2	I can learn <topic> if I get some help from the teacher	.56**	1	.61**	.67**	.74**	.53**	.64**	.67**	.57**
3	I can learn <topic> without any help from the teacher at all	.39**	.55**	1	.52**	.73**	.78**	.45**	.65**	.69**
4	I can solve all the easy tasks	.34**	.29*	.18	1	.79**	.51**	.82**	.72**	.49**
5	I can solve all the medium difficulty tasks	.30*	.61**	.59**	.56**	1	.79**	.77**	.79**	.64**
6	I can solve all the hard tasks	.23	.50**	.64**	.26*	.86**	1	.56**	.72**	.74**
7	I can persevere if I am faced with slightly difficult tasks	.22	.32*	.20	.31*	.34**	.32**	1	.87**	.57**
8	I can persevere if I am faced with medium difficulty tasks	.22	.44**	.41**	.31*	.61**	.62**	.85**	1	.82**
9	I can persevere if I am faced with very difficult tasks	.26*	.53**	.52**	.22	.70**	.77**	.58**	.86**	1

**Correlation is significant at the 0.01 level (2-tailed), *correlation is significant at the 0.05 level (2-tailed).

Grade 6 estimates are displayed below the diagonal, grade 10 above

Table 43: Correlation matrix Self-efficacy, lesson 3

	Item	1	2	3	4	5	6	7	8	9
1	I can learn <topic> if I get a lot of help from the teacher	1	.86**	.39**	.76**	.61**	.46**	.70**	.63**	.47**
2	I can learn <topic> if I get some help from the teacher	.74**	1	.61**	.81**	.80**	.68**	.68**	.75**	.62**
3	I can learn <topic> without any help from the teacher at all	.23*	.41**	1	.57**	.78**	.85**	.50**	.68**	.70**
4	I can solve all the easy tasks	.30*	.34**	.26*	1	.83**	.67**	.80**	.76**	.58**
5	I can solve all the medium difficulty tasks	.29*	.47**	.43**	.58**	1	.90**	.72**	.87**	.79**
6	I can solve all the hard tasks	.20	.34**	.65**	.39**	.79**	1	.59**	.79**	.82**
7	I can persevere if I am faced with slightly difficult tasks	.34**	.29*	.17	.33**	.43**	.37**	1	.87**	.65**
8	I can persevere if I am faced with medium difficulty tasks	.28*	.43**	.38**	.37**	.60**	.55**	.84**	1	.90**
9	I can persevere if I am faced with very difficult tasks	.22	.35**	.49**	.25*	.60**	.71**	.57**	.78**	1

**Correlation is significant at the 0.01 level (2-tailed), *correlation is significant at the 0.05 level (2-tailed).

Grade 6 estimates are displayed below the diagonal, grade 10 above

Table 44: Correlation matrix Self-efficacy, lesson 4

	Item	1	2	3	4	5	6	7	8	9
1	I can learn <topic> if I get a lot of help from the teacher	1	.90**	.38**	.74**	.68**	.47**	.71**	.58**	.37**
2	I can learn <topic> if I get some help from the teacher	.67**	1	.61**	.76**	.80**	.65**	.73**	.72**	.52**
3	I can learn <topic> without any help from the teacher at all	.36**	.56**	1	.58**	.78**	.87**	.51**	.67**	.63**
4	I can solve all the easy tasks	.38**	.40**	.37**	1	.86**	.63**	.87**	.74**	.55**
5	I can solve all the medium difficulty tasks	.34**	.45**	.39**	.46**	1	.87**	.79**	.83**	.68**
6	I can solve all the hard tasks	.40**	.43**	.53**	.23	.77**	1	.62**	.76**	.71**
7	I can persevere if I am faced with slightly difficult tasks	.58**	.51**	.20	.59**	.46**	.37**	1	.86**	.66**
8	I can persevere if I am faced with medium difficulty tasks	.43**	.53**	.20	.04	.76**	.71**	.71**	1	.88**
9	I can persevere if I am faced with very difficult tasks	.45**	.45**	.44**	.14	.75**	.84**	.50**	.82**	1

**Correlation is significant at the 0.01 level (2-tailed), *correlation is significant at the 0.05 level (2-tailed).

Grade 6 estimates are displayed below the diagonal, grade 10 above

4.4.2.2 Factor structure of my self-efficacy measure

Tables 45 to 49 display the fit indices for the measurement models I specified, for each of the five time points. For each time point, six alternative models were specified (see Figures 24 to 29 in section 4.3.6.2), and assessed in regard to fit indices and parameter estimates. The below tables indicate that the Unidimensional Model, Model 0, is a poor fit to the data, considering the fit indices (fit indices and cut-offs were presented in section 4.3.6). Fit indices for the Latent Facets Model, Model 1, ranges from "acceptable" to "good", but for each time point, this model results in negative residual variance on several items. According to the chosen cut-offs, Model 0 and Model 1 are not "accepted". According to my chosen cut-offs, Models 2, 3, 4, and 5 are all "accepted" considering fit indices and parameter estimates, and were thus compared in order to find the "best-fitting" model.

The Latent Levels Model (Model 2) is the most parsimonious, against which I compared the three Bi-factor Models (e.g. the differences in fit needed to be larger than the specified cut-offs in order for the more parsimonious Model 2 to be rejected). At pre-observations and lesson 2, the Latent Levels Model (Model 2) is equivalent in fit considering $\Delta\chi^2$, ΔCFI, and ΔSRMR, and thus the preferred model. Considering $\Delta\chi^2$ and ΔCFI the Bi-factor Easy Model (Model 3) is better fitting at lesson 1, while the Bi-factor Hard Model (Model 6) is better fitting at lesson 3. At lesson 4 Bi-factor Hard Model (Model 6) is better fitting considering $\Delta\chi^2$ and ΔSRMR, while the Latent Levels Model is equivalent (e.g. better fitting) considering ΔCFI. Considering CFI only, the Latent Levels Model is thus equivalent or better fitting at three time points, while the Bi-factor Easy and Hard Models are better fitting at one time point each. The ΔSRMR reaches the cut-off at a single time point only (lesson 4), supporting Model 2 as best fitting at all other time points. Considering the delta fit indices across all four lessons, the Latent Levels Model is overall better fitting. The alternative Bi-factor Models are

not consistent (the same Bi-factor Model is not better fitting at every time point), thus I cannot justify rejecting the more parsimonious Latent Levels Model for one of these. As such, I argue the Latent Levels Model is the best-fitting model for the SEGD considering fit indices as well as substantive considerations (e.g. within-bounds parameter estimates).

Table 45: Fit indices from confirmatory factor analyses, Self-Efficacy pre-obs.

		Fit indices						*Model comparison**			
	Model	χ^2	*p*	RMSEA	CFI	TLI	SRMR	ΔCFI	ΔSRMR	$\Delta\chi^2_{(\Delta df)}$	*p*
0	Unidimensional model	$\chi^2_{(90)}$= 721.09 χ^2/df= 8.01	*p*<.001	0.20	0.56	0.49	0.10				
1	Latent Facets model, correlated uniquenesses for levels**	$\chi^2_{(50)}$= 116.59 χ^2/df= 2.33	*p*<.001	0.09	0.95	0.90	0.08				
2	Latent Levels model, correlated uniquenesses for facets	$\chi^2_{(72)}$= 101.10 χ^2/df= 1.40	p=.014	0.05	0.98	0.97	0.06				
3	Bi-factor Easy model, correlated uniquenesses for facets	$\chi^2_{(64)}$= 90.19 χ^2/df= 1.41	p=.017	0.05	0.98	0.97	0.04	0.002	-0.017	$\Delta\chi^2_{(8)}$= -10.91	*p>.05*
4	Bi-factor Medium model, correlated uniquenesses for facets	$\chi^2_{(64)}$= 94.08 χ^2/df= 1.47	p=.009	0.05	0.98	0.97	0.05	0.002	-0.010	$\Delta\chi^2_{(8)}$= -7.02	*p>.05*
5	Bi-factor Hard model, correlated uniquenesses for facets	$\chi^2_{(64)}$= 94.27 χ^2/df= 1.47	p=.008	0.05	0.98	0.97	0.05	-0.001	-0.006	$\Delta\chi^2_{(8)}$= -6.83	*p>.05*

**Negative residual variance on several items, *Models 3-5 were compared against Model 2

Table 46: Fit indices from confirmatory factor analyses, Self-Efficacy lesson 1

		Fit indices						*Model comparison**			
	Model	χ^2	*p*	RMSEA	CFI	TLI	SRMR	ΔCFI	ΔSRMR	$\Delta\chi^2_{(\Delta df)}$	*p*
0	Unidimensional model	$\chi^2_{(27)}$= 207.93 χ^2/df= 7.70	*p*<.001	0.20	0.77	0.69	0.07				
1	Latent Facets model, correlated uniquenesses for levels**	$\chi^2_{(15)}$= 27.49 χ^2/df= 1.83	*p*=.025	0.07	0.98	0.96	0.06				
2	Latent Levels model, correlated uniquenesses for facets	$\chi^2_{(15)}$= 27.37 χ^2/df= 1.82	*p*=.026	0.07	0.98	0.96	0.03				
3	Bi-factor Easy model, correlated uniquenesses for facets	$\chi^2_{(11)}$= 7.03 χ^2/df= 0.64	*p*=.797	0.00	1.00	1.02	0.01	0.016	-0.021	$\Delta\chi^2_{(4)}$= -20.34	*p<.05*
4	Bi-factor Medium model, correlated uniquenesses for facets	$\chi^2_{(11)}$= 24.34 χ^2/df= 2.21	*p*=.011	0.09	0.98	0.94	0.03	-0.001	-0.003	$\Delta\chi^2_{(4)}$= -3.03	*p>.05*
5	Bi-factor Hard model, correlated uniquenesses for facets	$\chi^2_{(11)}$= 15.29 χ^2/df= 1.39	*p*=.17	0.05	0.99	0.98	0.02	0.010	-0.008	$\Delta\chi^2_{(4)}$= -12.08	*p<.05*

**Negative residual variance on several items, *Models 3-5 were compared against Model 2

Table 47: Fit indices from confirmatory factor analyses, Self-Efficacy lesson 2

		Fit indices						*Model comparison**			
	Model	χ^2	*p*	RMSEA	CFI	TLI	SRMR	ΔCFI	ΔSRMR	$\Delta\chi^2_{(\Delta df)}$	*p*
0	Unidimensional model	$\chi^2_{(27)}$= 379.11 χ^2/df= 14.04	*p*<.001	0.30	0.50	0.33	0.09				
1	Latent Facets model, correlated uniquenesses for levels**	$\chi^2_{(15)}$= 36.39 χ^2/df= 2.43	*p*=.002	0.97	0.97	0.93	0.09				
2	Latent Levels model, correlated uniquenesses for facets	$\chi^2_{(15)}$= 23.83 χ^2/df= 1.59	*p*=.068	0.06	0.99	0.97	0.03				
3	Bi-factor Easy model, correlated uniquenesses for facets	$\chi^2_{(11)}$= 21.94 χ^2/df= 1.99	*p*=.025	0.08	0.98	0.95	0.02	-0.003	-0.008	$\Delta\chi^2_{(4)}$= -1.89	*p>.05*
4	Bi-factor Medium model, correlated uniquenesses for facets	$\chi^2_{(11)}$= 24.08 χ^2/df= 2.19	*p*=.012	0.09	0.98	0.94	0.02	-0.006	-0.004	$\Delta\chi^2_{(4)}$= 0.25	*p>.05*
5	Bi-factor Hard model, correlated uniquenesses for facets	$\chi^2_{(11)}$= 18.40 χ^2/df= 1.67	*p*=.073	0.07	0.99	0.97	0.02	0.002	-0.005	$\Delta\chi^2_{(4)}$= -5.43	*p>.05*

**Negative residual variance on several items, *Models 3-5 were compared against Model 2

Table 48: Fit indices from confirmatory factor analyses, Self-Efficacy lesson 3

		Fit indices						*Model comparison**			
	Model	χ^2	*p*	RMSEA	CFI	TLI	SRMR	ΔCFI	ΔSRMR	$\Delta\chi^2_{(\Delta df)}$	*p*
0	Unidimensional model	$\chi^2_{(27)}$= 334.84 χ^2/df= 12.40	*p*<.001	0.27	0.58	0.44	0.09				
1	Latent Facets model, correlated uniquenesses for levels**	$\chi^2_{(15)}$= 66.09 χ^2/df= 4.41	*p*<.001	0.15	0.93	0.83	0.09				
2	Latent Levels model, correlated uniquenesses for facets	$\chi^2_{(15)}$= 32.90 χ^2/df= 2.19	*p*=.005	0.09	0.98	0.94	0.03				
3	Bi-factor Easy model, correlated uniquenesses for facets	$\chi^2_{(11)}$= 31.25 χ^2/df= 2.84	*p*=.001	0.11	0.97	0.91	0.02	-0.004	-0.010	$\Delta\chi^2_{(4)}$= -1.65	*p>.05*
4	Bi-factor Medium model, correlated uniquenesses for facets	$\chi^2_{(11)}$= 30.62 χ^2/df= 2.78	*p*=.013	0.11	0.97	0.91	0.03	-0.003	-0.005	$\Delta\chi^2_{(4)}$= -2.28	*p>.05*
5	Bi-factor Hard model, correlated uniquenesses for facets	$\chi^2_{(11)}$= 19.75 χ^2/df=1.80	*p*=.049	0.07	0.99	0.96	0.02	0.012	-0.013	$\Delta\chi^2_{(4)}$= -13.15	*p<.05*

**Negative residual variance on several items, *Models 3-5 were compared against Model 2

Table 49: Fit indices from confirmatory factor analyses, lesson 4

		Fit indices						*Model comparison**			
	Model	χ^2	*p*	RMSEA	CFI	TLI	SRMR	ΔCFI	ΔSRMR	$\Delta\chi^2_{(\Delta df)}$	*p*
0	Unidimensional model	$\chi^2_{(27)}$= 270.81 χ^2/df= 10.03	*p*<.001	0.27	0.66	0.55	0.08				
1	Latent Facets model, correlated uniquenesses for levels**	$\chi^2_{(15)}$= 49.61 χ^2/df= 3.31	*p*<.001	0.13	0.95	0.88	0.10				
2	Latent Levels model, correlated uniquenesses for facets	$\chi^2_{(15)}$= 22.46 χ^2/df= 1.50	*p*=.096	0.06	0.99	0.98	0.05				
3	Bi-factor Easy model, correlated uniquenesses for facets	$\chi^2_{(11)}$= 14.89 χ^2/df= 1.35	p=.188	0.05	1.00	0.98	0.02	0.005	-0.028	$\Delta\chi^2_{(4)}$= -7.57	*p>.05*
4	Bi-factor Medium model, correlated uniquenesses for facets	$\chi^2_{(11)}$= 14.52 χ^2/df= 1.32	p=.205	0.05	1.00	0.98	0.03	0.005	-0.019	$\Delta\chi^2_{(4)}$= -7.94	*p>.05*
5	Bi-factor Hard model, correlated uniquenesses for facets	$\chi^2_{(11)}$= 4.52 χ^2/df= 0.41	p=.952	0.00	1.00	1.03	0.02	0.010	-0.030	$\Delta\chi^2_{(4)}$= -17.94	*p<.05*

**Negative residual variance on several items, *Models 3-5 were compared against Model 2

The choice of a measurement model also includes aspects of pragmatism and coherence. The Bi-factor Models provide parameter estimates for each of the easy, medium, and hard level tasks, dis-attenuated for the other levels of difficulty (as well as facet specificity), in three separate models. In contrast to this, the Latent Levels Model provides parameter estimates for self-efficacy for easy, medium, and hard level tasks, dis-attenuated for facet specificity. As such the Latent Levels Model enables investigation into self-efficacy for different levels of difficulty, and provides a practical structure for easy, medium, and hard difficulty parcels in further analyses. Furthermore, a comparable Latent Levels Model was chosen as the best fitting model in Study 1 of the present thesis. When I choose Model 2, the Latent Levels Model, as my preferred measurement model, this is in view of fit indices, substantive interpretation, consistency with a previous study, and pragmatic considerations. For purposes of illustration I present in Figure 30 the parameter estimates from the Latent Levels Model for lesson 2.

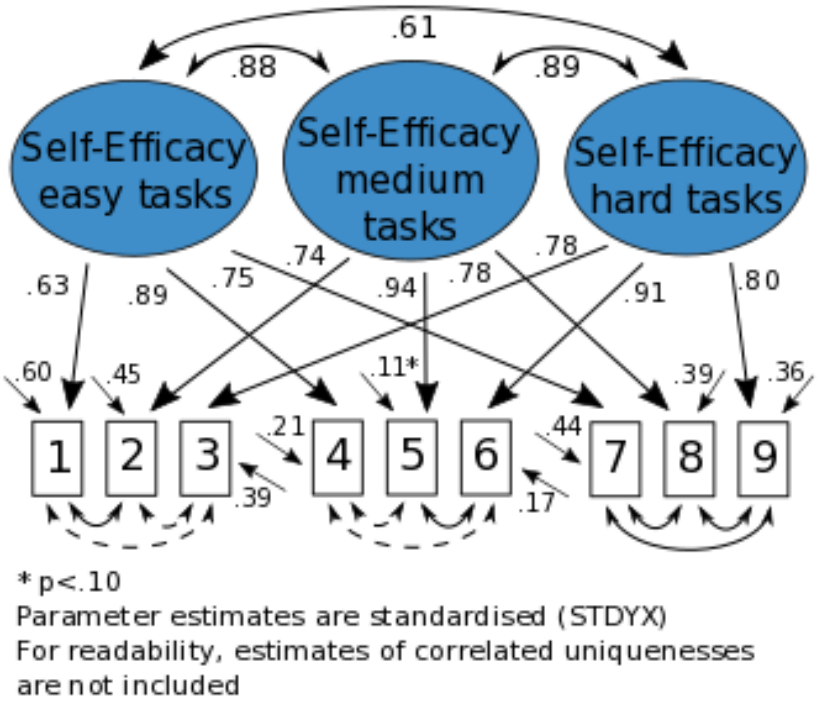


Figure 30: Latent Levels model, lesson 2

Having established a factor structure that fit across all the time points, I have also established configural invariance on my self-efficacy measure. The preferred measurement structure included three latent level constructs, which warrants investigating the correlations between these level constructs (e.g. the factors in the measurement model at each time point) within as well as across the time points. In Table 37, section 4.4.1.1 I presented the correlation coefficients for each of the three latent level constructs, at each time point. Within each time-point, the correlation estimates are consistently higher with constructs more closely associated in regard to level of difficulty (e.g. within each time-point, the easy level construct is more strongly correlated with the medium, than with the hard level construct).

Looking at the autocorrelations for each of the level constructs across time, a simplex model (Little, 2013, p. 184) is likely to fit the data well, particularly considering the grade 6 pattern of correlations. In grade 6, the autocorrelations consistently decrease across increasing time-spans, with some exceptions: The lesson 4 measures have higher estimates in relation to some of the pre-observation or lesson 1 measures, than preceding time points (e.g. lesson 3 measures). In the grade 10 data, the correlation estimates are generally higher (average estimate of .73, versus .55 for the grade 6 sample, as discussed in section 4.4.1.1), and also more stable as compared with the younger students. While some of the autocorrelations decrease with increasing time-spans (e.g. medium level construct at pre-observation time-point), some remain relatively stable, while others again increase (e.g. lesson 1 easy level construct), in absolute terms. I will investigate the simplex model in section 4.4.3.1. In the next section I will discuss weak and strong invariance of my measure across the four lessons.

4.4.2.3 Measurement invariance of my self-efficacy measure

In order to test whether the constructs were measurement invariant across the four lessons, I specified models in which relevant factor loadings were constrained to be equal across time,

and compared these with less restrained models. I used my preferred measurement model from section 4.4.2.2, however when I added the four time points this model resulted in several error messages. As is evident when investigating the correlation matrix in Table 37 the three latent level constructs are highly correlated. In order to avoid issues with co-linearity, I proceeded to model each of the self-efficacy level constructs separately (see Figures 31 to 33). I was interested primarily in whether the factor loadings and item intercepts stayed invariant over time (I saw invariant residual variances, e.g. strict measurement invariance, as too strong a requirement, as argued by Little, 2013, p. 143), thus I investigated weak and strong measurement invariance for each of the three latent level factors. Table 50 presents the fit indices for models with increasing equality constraints applied, the difference in CFI, and the $\chi 2$ difference in relation to the differences in degrees of freedom.

Table 50: Fit indices Self-Efficacy invariance testing

		Fit indices						*Model comparison***			
	Model	χ^2	*p*	RMSEA	CFI	TLI	SRMR	ΔCFI	$\Delta\chi^2_{(\Delta df)}$	*p*	*pass?*
0	Latent levels model, CU's for facets, no constraints**	$\chi^2_{(438)}$= 957.13	*p*<.001	0.08	0.91	0.87	0.06				
1	Latent easy level model, no constraints	$\chi^2_{(30)}$= 51.24	p=.009	0.06	0.96	0.92	0.04				
1b	Latent easy level model, weak invariance	$\chi^2_{(36)}$= 50.88	p=.051	0.05	0.98	0.95	0.06	0.011	$\Delta\chi^2_{(6)}$= -0.36	*p>.05*	*yes*
1c	Latent easy level model, strong invariance	$\chi^2_{(45)}$= 73.61	p=.005	0.06	0.95	0.93	0.14	-0.024	$\Delta\chi^2_{(9)}$= 22.73	*p*<.01	*no*
1d	Latent easy level model, *partial* strong invariance (*item 7, lesson 4)	$\chi^2_{(44)}$= 46.07	p=.387	0.02	1.00	1.00	0.08	0.021	$\Delta\chi^2_{(8)}$= -4.81	*p>.05*	*yes*
1e	Latent easy level model, weak invariance (grade 10 only)	$\chi^2_{(36)}$= 50.07	p=.06	0.06	0.97	0.95	0.13				
1f	Latent easy level model, strong invariance (grade 10 only)	$\chi^2_{(45)}$= 51.11	p=.246	0.04	0.99	0.98	0.12	0.015	$\Delta\chi^2_{(9)}$= -1.04	*p>.05*	*yes*
2	Latent medium difficulty level model, no constraints	$\chi^2_{(30)}$= 39.41	p=.117	0.04	0.99	0.98	0.02				
2b	Latent medium difficulty level model, weak invariance	$\chi^2_{(36)}$= 43.19	p=.191	0.03	0.99	0.99	0.04	0.002	$\Delta\chi^2_{(6)}$= 3.78	*p>.05*	*yes*
2c	Latent medium difficulty level model, strong invariance	$\chi^2_{(45)}$= 52.75	p=.200	0.03	0.99	0.99	0.05	-0.001	$\Delta\chi^2_{(9)}$= 9.56	*p>.05*	*yes*
3	Latent hard level model, no constraints	$\chi^2_{(30)}$= 20.91	p=.891	0.00	1.00	1.02	0.02				
3b	Latent hard level model, weak invariance	$\chi^2_{(36)}$= 26.89	p=.864	0.00	1.00	1.02	0.03	0.000	$\Delta\chi^2_{(6)}$= 5.98	*p>.05*	*yes*
3c	Latent hard level model, strong invariance	$\chi^2_{(45)}$= 31.97	p=.930	0.00	1.00	1.02	0.02	0.000	$\Delta\chi^2_{(9)}$= 5.08	*p>.05*	*yes*

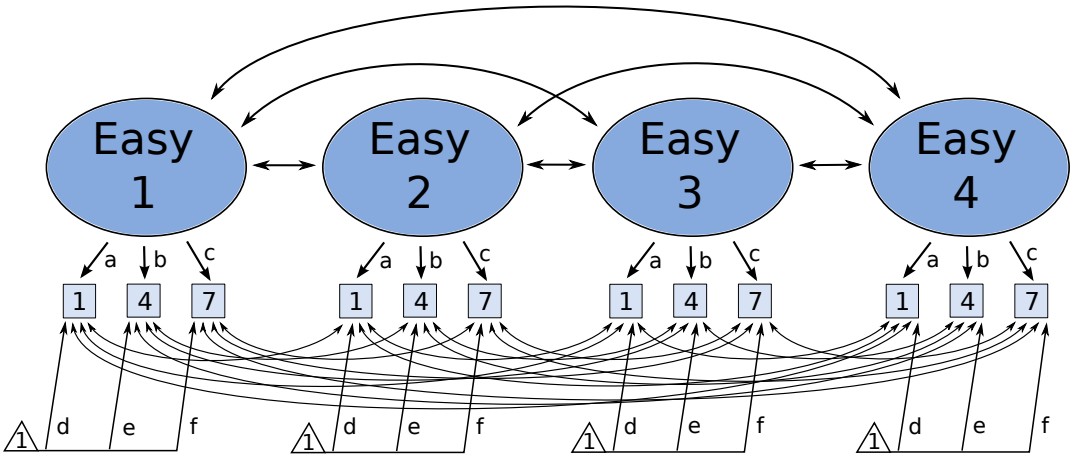
* *Item not constrained as equal*

***The Latent Levels Model resulted in an error message, thus each of the latent level structures were investigated separately*

****Each model was compared with the equivalent model with one less constraint*

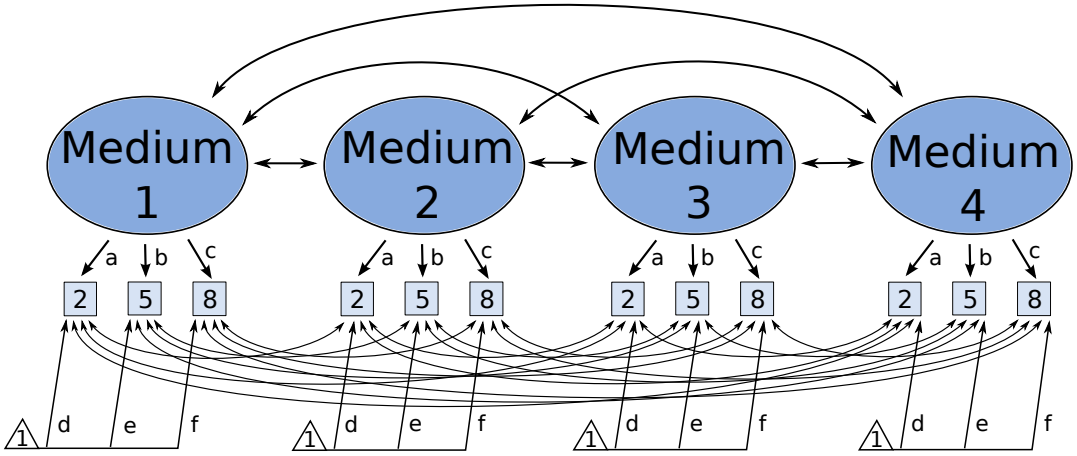
The models for all three latent level factors proved to be factorially invariant over time (as evident in the CFI and $\chi 2$ difference scores). However, only the medium difficulty and the hard level factor structure proved to be strongly invariant, that is, with invariant item intercepts across the four time points. I proceeded to test for strong measurement invariance for the easy level construct for each of the grade levels separately (as well as in a two-group model). The model did not converge (due to a small sample size) when including only the grade 6 sample. The small grade 6 sample size was also the reason for an error message when I specified a two-group model. However, the model converged for the older sample, and nested models demonstrated the easy level self-efficacy construct was strongly invariant in regard to the grade 10 sample. This indicates the problem of non-invariance is in relation to the grade 6 sample.

The modification indices indicated the problem is in relation to item 7 (*I can persevere- when faced with somewhat hard tasks*), at lesson 4. This indicates the meaning of this item may have changed for grade 6 students across the lessons (considering it was the last of the lessons in the sequence). According to Little (2013) it is still reasonable to investigate changes in a construct over time so long as a majority (e.g. 2 out of 3) of the indicators for the construct are strongly invariant, which is the case with the present construct. Furthermore, "partial invariance is not a problem if there aren't too many loadings or intercepts that change meaning over time" (Little, 2013, p. 159). Accordingly, I specified a partially strong invariant model for the easy level latent construct, where the intercept of item 7 (parameter *f* in Figure 31) at lesson 4 was not constrained to be equal with the other lessons. The model with partial strong invariance was supported, considering the fit indices. The model specifications for the strongly invariant measurement models are displayed in Figures 31 to 33. Note that in the case of the easy level construct a partially strong invariant model was supported.



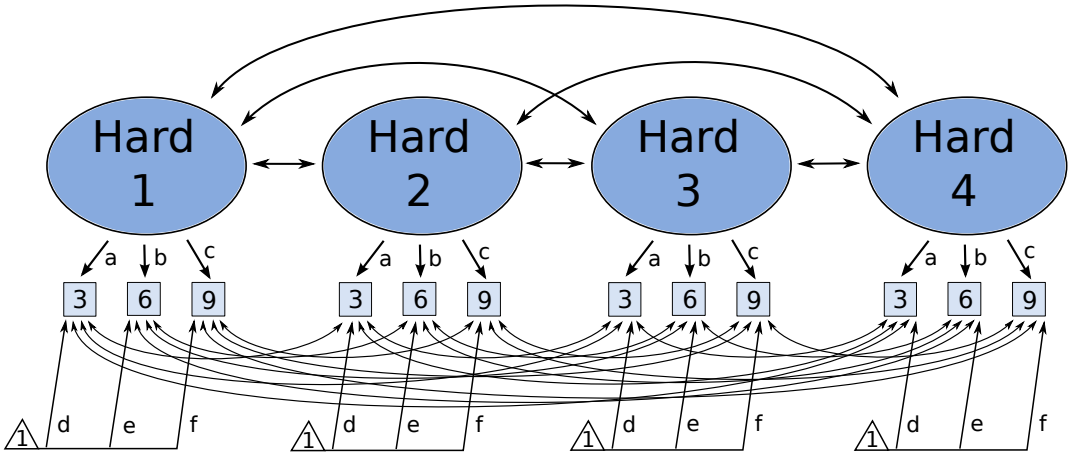
Easy 1 - 4 = Students' self-efficacy for easy tasks
a, b, and c are factor loadings; d, e, and f are manifest (item) means
Weak invariance: parameters a, b, and c are constrained as equal
Strong invariance: parameters d, e, and f are also constrained as equal

Figure 31: Invariance model self-efficacy, easy level construct



Medium 1 - 4 = Students' self-efficacy for medium difficulty tasks
a, b, and c are factor loadings; d, e, and f are manifest (item) means
Weak invariance: parameters a, b, and c are constrained as equal
Strong invariance: parameters d, e, and f are also constrained as equal

Figure 32: Invariance model self-efficacy, medium difficulty level construct



Hard 1 - 4 = Students' self-efficacy for hard tasks
a, b, and c are factor loadings; d, e, and f are manifest (item) means
Weak invariance: parameters a, b, and c are constrained as equal
Strong invariance: parameters d, e, and f are constrained as equal

Figure 33: Invariance model self-efficacy, hard level construct

4.4.2.4 Summary of results for RQ1

On the basis of comparing a series of theoretically proposed models, I retained the factor structure from my previous study (see section 2.5.1, study 1), including three latent self-efficacy level constructs, with correlated uniquenesses for facets. I established configural and weak invariance, and found that my measures were strong factorially invariant across time in the case of the medium and hard self-efficacy constructs. In regard to the easy self-efficacy construct my measure was partially strong factorially invariant. This was due to a single item (*I can persevere – when faced with somewhat hard tasks*), which for the grade 6 sample could not be constrained as equivalent in intercept at lesson 4, as the other lessons. This indicates that in the case of the grade 6 students, this item may have changed meaning over the course of the lessons. It is possible that the experiences from the mathematics lessons informed the students' growing conceptualisation of what "perseverance" in mathematics means, or how difficult "somewhat hard tasks" are (NB! The English term perseverance is a translation from the Norwegian term used in the questionnaire. The original term, directly translated, is closer

to "Not give up", and may have higher face validity for young students than the term "perseverance"). This illustrates one challenge of measuring students' motivation processes, where I am trying to dis-attenuate change due to learning or cognitive maturation from "actual" change in the construct of interest. On the basis of the testing of my factor structure, the invariance testing provided satisfying evidence that my measure stayed invariant over time, and can be used to investigate changes in stability and mean self-efficacy scores across lessons in mathematics.

4.4.3 RQ2: Stability of students' rank-order

4.4.3.1 Autoregressive path models for a simplex change process

In order to investigate to what degree the rank-order of students' self-efficacy remained stable over time, I specified autoregressive path models for each of the three self-efficacy level constructs. Based on the correlation matrix for the latent level constructs presented in Table 37 and the results from the invariance testing in section 4.4.2.3 I specified simplex models of change for each of the latent level constructs (see Table 51). In these models students' self-efficacy at time point *t* predicted their self-efficacy at time point *t+1*. Autoregressive models based on the simplex structure resulted in very good model fit indices for all three level constructs (see Models 1, 2, and 3 in Table 51).

Table 51: Fit indices Self-Efficacy simplex models

		Fit indices						*Model comparison**			
	Model	χ^2	*p*	RMSEA	CFI	TLI	SRMR	ΔCFI	$\Delta\chi^2_{(\Delta df)}$	*p*	*pass?*
1	Latent easy level model, *partial* strong invariance, simplex	$\chi^2_{(47)}$= 51.47	p=.303	0.02	0.99	0.99	0.08				
1b	Model 1, plus equality constraints on all stability paths	$\chi^2_{(49)}$= 50.59	p=.411	0.01	1.00	1.00	0.09	0.005	$\Delta\chi^2_{(2)}$= -.88	p>.05	*yes*
1c	Model 1b, regressed on pre-observations and grade.	$\chi^2_{(117)}$= 154.27	p=.012	0.04	0.96	0.95	0.09				
2	Latent medium level model, strong invariance, simplex	$\chi^2_{(48)}$= 65.20	p=.050	0.05	0.98	0.97	0.05				
2b	Model 2, plus equality constraints on all stability paths	$\chi^2_{(50)}$= 65.20	p=.073	0.04	0.98	0.98	0.05	0.002	$\Delta\chi^2_{(2)}$= 0	p>.05	*yes*
2c	Model 2b, regressed on pre-observations and grade.	$\chi^2_{(118)}$= 160.67	p=.006	0.05	0.97	0.97	0.06				
3	Latent hard level model, strong invariance, simplex	$\chi^2_{(48)}$= 42.08	p=.713	0.00	1.00	1.01	0.04				
3b	Model 3, plus equality constraints on all stability paths	$\chi^2_{(50)}$= 43.37	p=.735	0.00	1.00	1.01	0.04	0.000	$\Delta\chi^2_{(2)}$= 1.29	p>.05	*yes*
3c	Model 3b, regressed on pre-observations and grade.	$\chi^2_{(118)}$= 165.81	p=.002	0.05	0.97	0.97	0.07				

**Models 1b, 2b, and 3b are compared against Models 1, 2, and 3, respectively*

Next, I included equality constraints on the autoregressive paths in order to investigate stability over time (Models 1b, 2b, and 3b in Table 51). There was no significant detriment in model fit through imposing equality constraints on either of the between-lesson stability paths. This indicates there were no significant differences in the stability of the rank-order of students' self-efficacy across lessons, for either of the self-efficacy constructs. I proceeded to include grade level and students' mathematics self-efficacy (as measured at the pre-observation time point) as covariates in each of the simplex models for the latent self-efficacy constructs with equality constraints for stability (Models 1c, 2c, and 3c in Table 51).

The parameter estimates from the autoregressive models including covariates are shown in Figures 34 to 36. In study 4, section 5.4.2.3 I will display the unstandardised autoregressive paths for self-efficacy, thus I choose to display the standardised stability paths here. The autoregressive paths demonstrate that for each of the level constructs, the stability of the rank-order of self-efficacy scores was consistently predicted by the previous time point. In the case of the easy self-efficacy construct this stability ranged from .85 to .98 (unstandardised estimate .94 for all lesson time lags), in the case of the medium self-efficacy construct it ranged from .90 to .98 (unstandardised estimate .96), and in the case of the hard self-efficacy construct the stability paths ranged from .88 to .91 (unstandardised estimate .92). Across all three self-efficacy constructs, the lowest stability score was related to the path from the first to the second lesson, however there were no statistical differences between the stability paths, for either of the level constructs. While the overall stability estimates appeared somewhat higher in the case of the easy and medium difficulty, as compared to the hard self-efficacy construct, these differences were not tested significantly. Furthermore, in the case of all self-efficacy level constructs, students' mathematics self-efficacy (at pre-observations) predicted students' lesson-specific self-efficacy at the beginning of lesson 1. In addition, grade

significantly predicted students' self-efficacy, where a higher grade (e.g. grade 10) predicted lower self-efficacy scores at the beginning of lesson 1. Note that Figures 34 to 36 are based on models with invariance constraints (see section 4.4.2.3), and that estimates of correlated uniquenesses are not displayed for the sake of readability.

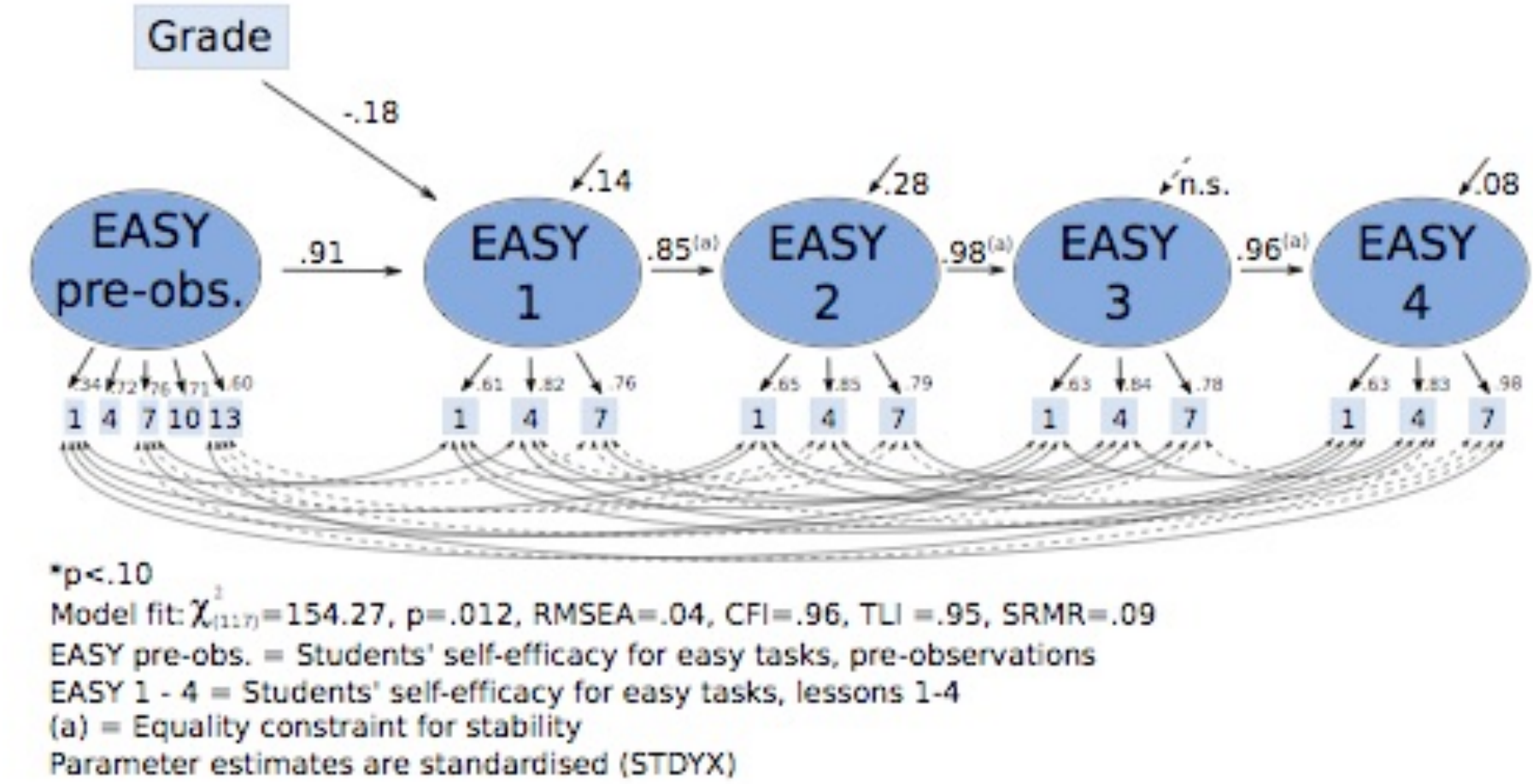


Figure 34: Autoregressive model Self-Efficacy, easy tasks

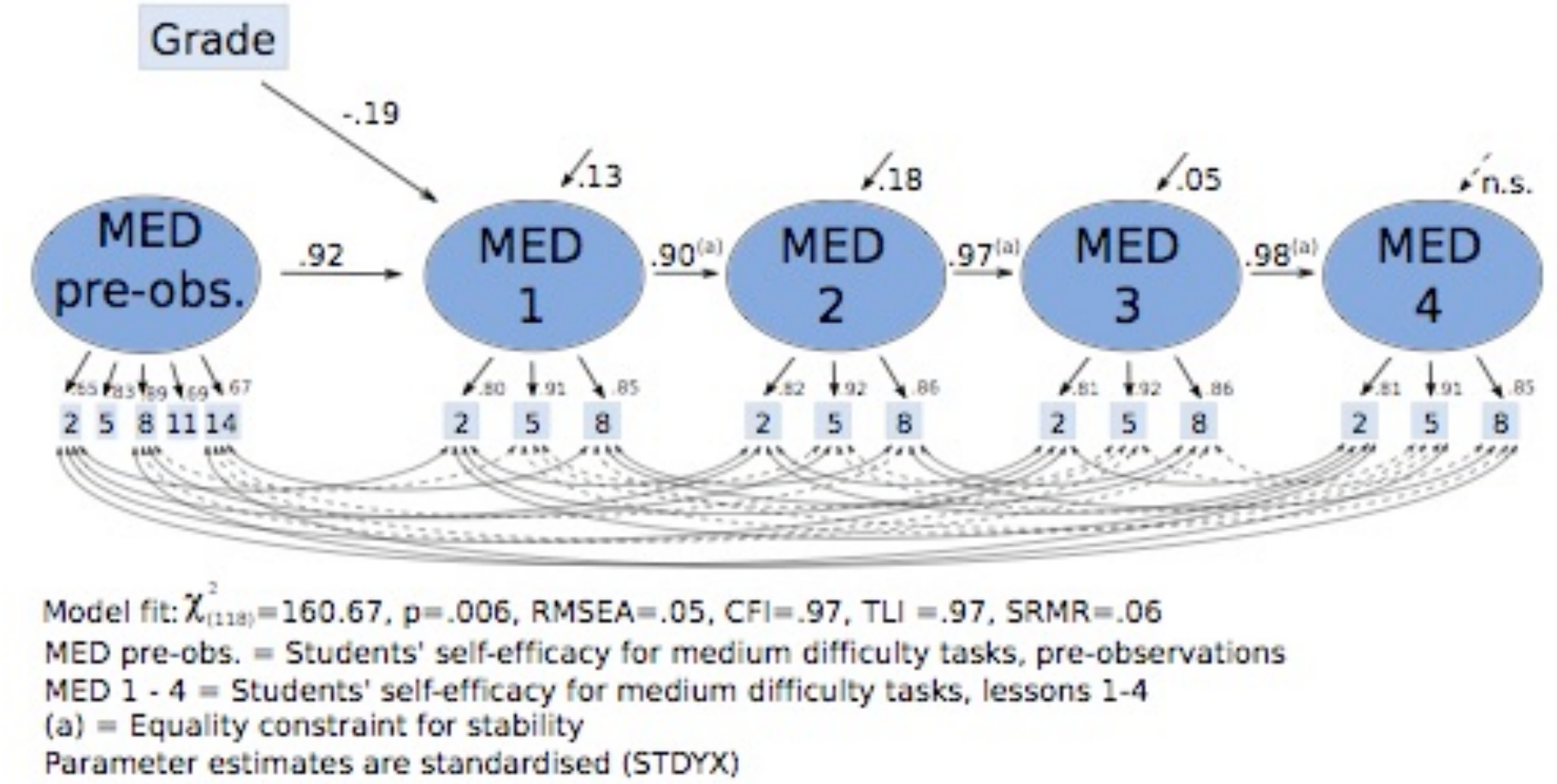


Figure 35: Autoregressive model Self-Efficacy, medium difficulty tasks

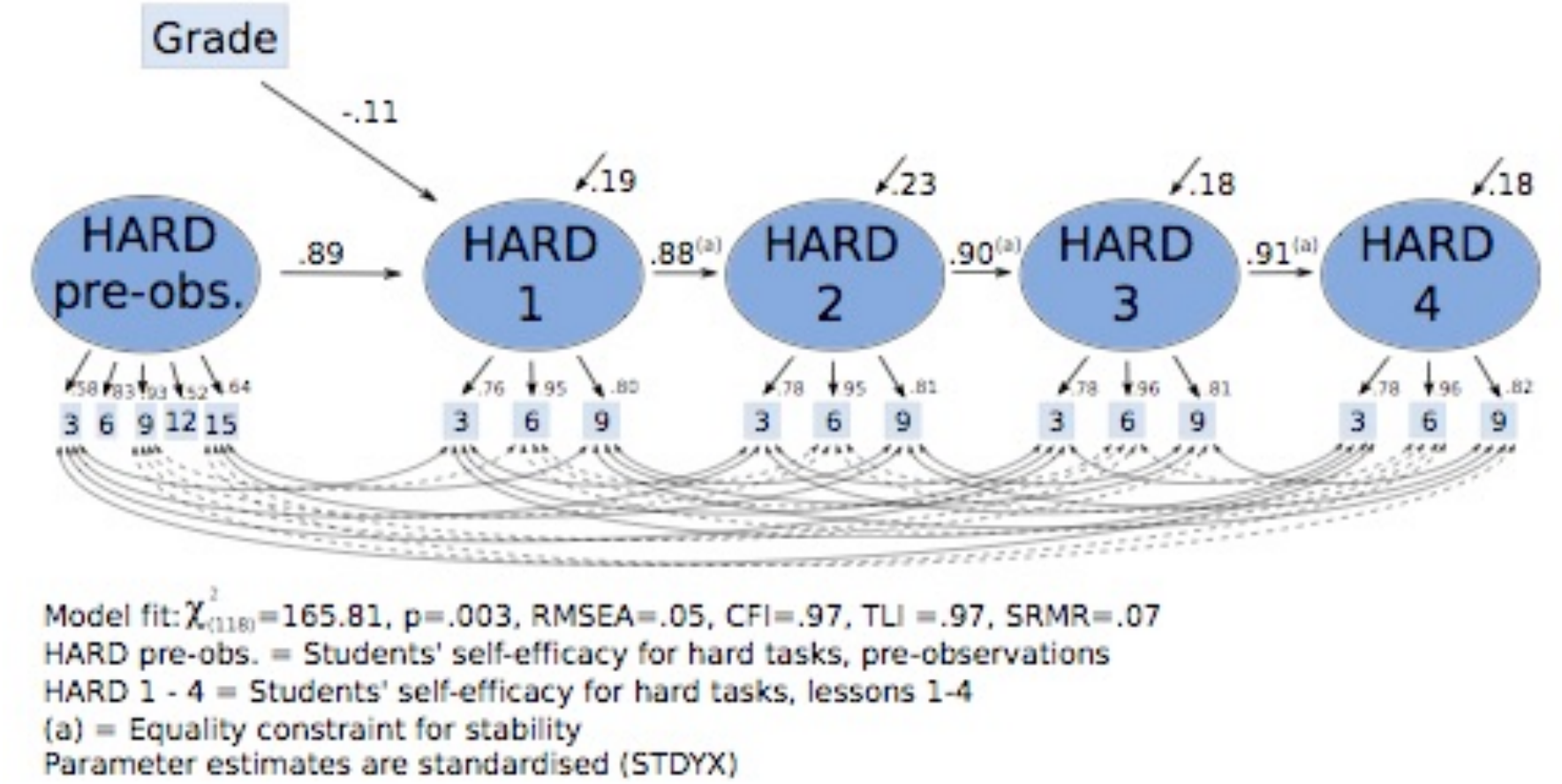


Figure 36: Autoregressive model Self-Efficacy, hard tasks

4.4.3.2 *Summary of results for RQ2*

For all three self-efficacy constructs, students' rank-order remained highly stable, with standardised paths ranging from .85 to .98. While the absolute stability parameters were consistently lower between the earlier, as compared with the later, lessons in the sequence, these differences did not prove to be statistically significant. There were only small differences in the stability parameters for self-efficacy for different levels of task difficulty, and as these were specified in separate models they were not investigated statistically. The simplex model fit well for all three self-efficacy constructs, which supports the assumption that self-efficacy change is a gradual and stable process, and indicates rank order changes in self-efficacy were not influenced by (non-measured) contextual factors to a large degree. This assumption will be further investigated in study 4, where measures of the theoretically strongest source of self-efficacy, students' mastery experiences, are included in the analyses.

The autoregressive path models demonstrated grade level predicted students' self-efficacy scores at the first lesson negatively, e.g. older students had lower self-efficacy at the beginning of lesson 1 than younger students. Students' self-efficacy for mathematics (a more general measure) at the pre-observation time point significantly predicted their lesson-specific mathematics self-efficacy (a more specific measure) at the beginning of the first lesson, in a similar magnitude as the stability path from the first to the second lesson.

4.4.4 RQ3: Self-efficacy change across lessons

Having investigated the rank-order stability of students' self-efficacy, I proceeded to investigate the mean structure of the self-efficacy constructs across a sequence of lessons.

4.4.4.1 Variance component models

Given the nestedness of my data (students nested within 9 classes), I applied variance component models to determine the portion of variance of students' self-efficacy scores at the class versus the individual level, e.g. the intra-class correlation coefficient (ICC). The intra-class correlation coefficient reflects the variability in students' self-efficacy predicted by students' class membership relative to the total variability in students' self-efficacy scores (e.g. class-level and individual-level variability combined). As a rule of thumb, between-level variability (here: variance between classes) is considered high if the ICC is higher than 5 %, depending on the context. I calculated the ICC at each time point, for each of my level constructs. The results from these analyses are presented in Tables 52 to 54.

For all three level constructs, across all time points, the class-level ("between") variance ranged from 0.3% to 18.6%. The lowest percentage of variance at the classroom level was in association with the second lesson, consistently for all three self-efficacy constructs. Considering all time points, the between-level variance was lower for the easy level self-efficacy construct, which ranged from 0.3% to 13.6%, followed by the medium difficulty level self-efficacy construct (range from 4.2% to 15.4%), and the hard level self-efficacy construct (range from 8.7% to 18.6%).

Table 52: Variance component models, Self-Efficacy easy level constructs

Model	Within var.	Between var.	ICC (Mplus)
Self-Efficacy Easy parcel, pre-observations. Within and between variance modelled	1.775	0.098	0.053
Self-Efficacy Easy parcel, lesson 1. Within and between variance modelled	1.956	0.309	0.137
Self-Efficacy Easy parcel, lesson 2. Within and between variance modelled	2.424	0.007	0.002
Self-Efficacy Easy parcel, lesson 3. Within and between variance modelled	2.448	0.182	0.070
Self-Efficacy Easy parcel, lesson 4. Within and between variance modelled	2.285	0.187	0.076

Table 53: Variance comp. models, Self-Efficacy medium difficulty level constructs

Model	Within var.	Between var.	ICC (Mplus)
Self-Efficacy Medium parcel, pre-observations. Within and between variance modelled	2.903	0.478	0.141
Self-Efficacy Medium parcel, lesson 1. Within and between variance modelled	3.514	0.689	0.164
Self-Efficacy Medium parcel, lesson 2. Within and between variance modelled	3.714	0.163	0.043
Self-Efficacy Medium parcel, lesson 3. Within and between variance modelled	3.817	0.694	0.154
Self-Efficacy Medium parcel, lesson 4. Within and between variance modelled	3.873	0.519	0.118

Table 54: Variance component models, Self-Efficacy hard level constructs

Model	Within var.	Between var.	ICC (Mplus)
Self-Efficacy Hard parcel, pre-observations. Within and between variance modelled	3.806	0.872	0.186
Self-Efficacy Hard parcel, lesson 1. Within and between variance modelled	5.430	0.900	0.142
Self-Efficacy Hard parcel, lesson 2. Within and between variance modelled	6.059	0.577	0.087
Self-Efficacy Hard parcel, lesson 3. Within and between variance modelled	6.174	1.035	0.143
Self-Efficacy Hard parcel, lesson 4. Within and between variance modelled	6.283	0.752	0.107

As demonstrated in Tables 52 to 54, only two of the ICC scores were lower than 5%, both at the lesson 2 time-point. Within- and between-variance can be handled with multilevel models or cluster analyses, however in my data there are too few groups at the between-level (only 9

classes) to employ either of these analytic approaches. It was not possible to specify models in which standard errors were adjusted by the clustering at the classroom level because there were too many model parameters of interest and too few level 2 units. As an alternative I specified dummy-variables for each of the classes, in order to investigate differences between the classes, in regard to students' self-efficacy. I first established which single class was closer to the average self-efficacy score, at each time point and for each latent self-efficacy construct. I proceeded to run simple models to investigate which classes differed significantly from this average class, displayed in Tables 55 to 57.

Table 55: Difference from average class, self-efficacy easy tasks

Grade	Class	Pre-obs	Lesson 1	Lesson 2	Lesson 3	Lesson 4
6	21	x	x	x	x	0.69* (.18)
6	23	x	x	x	0.56 (.14)	0.84 (.23)
6	24	x	x	x	x	x
6	31	x	x	NA	x	NA
10	10	x	x	x	x	x
10	11	-0.57 (-.24)	x	x	-0.99 (-.22)	x
10	20	x	x	x	x	x
10	35	x	x	x	x	NA
10	38	0.41* (.16)	1.06* (.24)	x	x	0.73* (.17)

Significant difference scores (STDYX in brackets)
x = does not differ significantly, *p<.10.

Table 56: Difference from average class, self-efficacy medium difficulty tasks

Grade	Class no	Pre-obs	Lesson 1	Lesson 2	Lesson 3	Lesson 4
6	21	x	x	x	x	x
6	23	x	1.00 (.18)	x	x	x
6	24	x	x	x	x	x
6	31	x	x	NA	-1.32 (-.17)	NA
10	10	x	x	x	-1.10* (-.20)	-1.03* (-.20)
10	11	-1.47 (-.32)	-1.41 (-.21)	x	-2.10 (-.36)	-1.36 (-.24)
10	20	x	-1.41 (-.24)	x	-1.24* (-.24)	-1.62 (-.33)
10	35	x	x	x	x	NA
10	38	1.14 (.23)	1.50 (.23)	x	x	x

Significant difference scores (STDYX in brackets)
x = does not differ significantly, *p<.10.

Table 57: Difference from average class, self-efficacy hard tasks

Grade	Class no	Pre-obs	Lesson 1	Lesson 2	Lesson 3	Lesson 4
6	21	x	x	x	x	1.29* (.21)
6	23	x	x	x	x	x
6	24	x	x	x	x	1.41* (.19)
6	31	x	x	NA	-1.80 (-.17)	NA
10	10	x	x	x	x	x
10	11	-1.96 (-.39)	-2.13 (-.28)	-2.63 (-.36)	-2.76 (-.34)	x
10	20	x	-1.41* (-.21)	x	x	x
10	35	x	x	x	x	NA
10	38	x	1.41 (.19)	x	x	1.19* (.17)

Significant difference scores (STDYX in brackets)
x = does not differ significantly, *p<.10.

Some tendencies are noteworthy: Significant difference scores within each class are consistent, as in either consistently higher or consistently lower than the average class, across all three levels and all five time points. There are more significant differences associated with self-efficacy for medium difficulty tasks, as compared with easy and hard tasks. As expected from the variance component models, there are less significant differences associated with lesson 2, while the highest number of significant differences is in relation to lesson 4. A single grade 10 class (class 11) accounts for a high proportion of the significant differences, with consistently lower average self-efficacy scores, as compared against the average class at each time point. Class 11 had relatively low mean self-efficacy and also low mean national test scores, as compared with most other classes, during phase 1 of this thesis (see Table 30, section 4.3.2.1). The data in the current study does not give any insights into the reasons for differences between the classes in terms of students' self-efficacy. It is likely some of the class-level variance is due to differences between the topics and/or the tasks the students engaged with, while some of the variance (e.g. in regard to class 11) may be related to more stable factors such as teacher, school, or home differences.

4.4.4.2 Latent growth models

In order to investigate mean level changes in students' self-efficacy across a sequence of lessons in mathematics, I used latent growth models. For each of the latent self-efficacy constructs, I specified a series of (nested) models, to allow comparison of model fits. I started with an unconditional model, in which I specified lesson 1 and lesson 4 at 0 and 1, respectively, allowing the in-between time points to be estimated freely. The unconditional model represents the best-fitting and least parsimonious model (Little, 2013), as the slope parameter is not "forced" to follow a particular shape; rather each slope parameter is freely estimated. I used the unconditional model as comparison model for a linear and a quadratic model, and differences in CFI and χ^2 for model comparison (see section 4.3.6). As discussed in section 4.3.6.2 I investigated mean level changes for the two groups separately. I present the parameter estimates for each of the models I tested (that resulted in meaningful parameter estimates), not just the best-fitting model. Some times the difference in fit between two competing models was not large (see e.g. Table 58 where the fit indices are "so good" overall there is little room for improvement, or for the alternative models to outcompete the more parsimonious linear model), meaning the choice of a "preferred model" could seem arbitrary. Thus, I reasoned it was meaningful to inspect whether or not there were major differences in estimated parameters as a function of the type of growth model selected. If not, this would raise my confidence that the final parameter estimates were a good estimation of the change process (at the group level). In latent growth curve modelling, the term "level" is generally used to refer to average scores at the zero-point in time (e.g. intercept), while in study 1 of this thesis I introduced the term "level" in referring to the degree of objective or perceived difficulty of tasks. For the present research question I shall avoid the use of level when referring to task difficulty, in order not to create confusion. I present the results according to each of the latent self-efficacy constructs, starting with students' self-efficacy for easy tasks.

Self-efficacy easy tasks

Tables 58 and 60 display the fit indices for the nested latent growth models for the easy self-efficacy construct. In the case of grade 6, all three models resulted in very good fit. Considering the chosen fit differences (see section 4.3.6), the more parsimonious, linear growth model (Model 2b) was supported as best-fitting. In the case of grade 10 the unconditional model (Model 1b) fit significantly better than the linear or quadratic models, and was supported as best-fitting. Tables 59 and 61 display the estimated parameters and model-explained variances for each of the growth models specified (models with out-of-bounds parameter estimates are not included), allowing for a comparison of intercept and slope parameters.

Table 58: Fit indices growth models easy tasks, grade 6

		Fit indices						*Model comparison*		
	Model	χ^2	*p*	RMSEA	CFI	TLI	SRMR	ΔCFI	$\Delta\chi^2_{(\Delta df)}$	*p*
1	Latent growth model with easy parcels, L1-4, unconditional*	$\chi^2_{(3)}$= 1.29	*p*=.731	0.00	1.00	1.06	0.08			
1b	Latent growth model with easy parcels, L1-4, unconditional. SEeasy_1 and SEeasy_4 > 0	$\chi^2_{(3)}$= 1.25	*p*=.742	0.00	1.00	1.06	0.17			
2	Latent growth model with easy parcels, L1-4, linear*	$\chi^2_{(5)}$= 4.27	*p*=.511	0.00	1.00	1.02	0.07			
2b	**Latent growth model with easy parcels, L1-4, linear. SEeasy_4 > 0**	**$\chi^2_{(5)}$= 3.79**	***p*=.579**	**0.00**	**1.00**	**1.03**	**0.14**	0.000	$\Delta\chi^2_{(2)}$= 2.54	*p>.05*
3	Latent growth model with easy parcels, L1-4, quadratic*	$\chi^2_{(1)}$= 0.06	*p*=.803	0.00	1.00	1.10	0.01			
3b	Latent growth model with easy parcels, L1-4, quadratic. Quad. var. @0, SEeasy_4 > 0.	$\chi^2_{(4)}$= 3.17	*p*=.530	0.00	1.00	1.02	0.15	0.000	$\Delta\chi^2_{(1)}$= 1.92	*p>.05*

*Negative residual variance of one item, **Negative residual variances of slope and quadratic factors

Table 59: Parameter estimates easy task models, grade 6

Unstandardized parameter estimates and model-explained variances

	Model	Fixed effects			Random effects			Variances explained				Covariances		
		Intercept	Slope	Quad.	Intercept	Slope	Quad.	L1	L2	L3	L4	I with S	I with Q	S with Q
1b	Latent growth model with easy parcels, L1-4, unconditional. SEeasy_1 and SEeasy_4 > 0	9.10	0.51	-	1.66	1.30	-	1.00	0.71	0.64	1.00	-1.06	-	-
2b	**Latent growth model with easy parcels, L1-4, linear. SEeasy_4 > 0**	9.14	0.16	-	1.41	0.10*	-	0.82	0.79	0.64	1.00	-0.25*	-	-
3b	Latent growth model with easy parcels, L1-4, quadratic. Quad. var. @0, SEeasy_4 > 0.	9.11	0.24	n.s.	1.41	0.10*	0**	0.82	0.79	0.64	1.00	-0.25*	-	-

*p<.10, ** variance fixed at 0

Table 60: Fit indices growth models easy tasks, grade 10

		Fit indices						*Model comparison*		
	Model	χ^2	*p*	RMSEA	CFI	TLI	SRMR	ΔCFI	$\Delta\chi^2_{(\Delta df)}$	*p*
1	Latent growth model with easy parcels, L1-4, unconditional*	$\chi^2_{(3)}$= 8.19	*p*=.042	0.14	0.96	0.93	0.03			
1b	**Latent growth model with easy parcels, L1-4, unconditional. SEeasy_2 > 0.**	**$\chi^2_{(3)}$= 3.64**	***p*=.304**	**0.05**	**1.00**	**0.99**	**0.03**			
2	Latent growth model with easy parcels, L1-4, linear**	$\chi^2_{(5)}$= 16.94	*p*=.005	0.16	0.92	0.90	0.34			
2b	Latent growth model with easy parcels, L1-4, linear. Slope var. @0	$\chi^2_{(7)}$= 19.45	*p*=.007	0.14	0.91	0.93	0.33	–0.082	$\Delta\chi^2_{(4)}$= 15.81	*p<.05*
3	Latent growth model with easy parcels, L1-4, quadratic***	$\chi^2_{(1)}$= .013	*p*=.909	0.00	1.00	1.04	0.00			
3b	Latent growth model with easy parcels, L1-4, quadratic. Slope & quad. var. @0	$\chi^2_{(6)}$= 18.81	*p*=.005	0.15	0.91	0.91	0.33	–0.084	$\Delta\chi^2_{(3)}$= 15.17	*p<.05*

*Negative residual variance of one item, **Negative residual variance of one factor, ***Negative residual variance of two items

Table 61: Parameter estimates easy task models, grade 10

Unstandardized parameter estimates and model-explained variances

	Model	Fixed effects			Random effects			Variances explained				Covariances		
		Intercept	Slope	Quad.	Intercept	Slope	Quad.	L1	L2	L3	L4	I with S	I with Q	S with Q
1b	**Latent growth model with easy parcels, L1-4, unconditional. SEeasy_2 > 0.**	8.82	n.s.	-	2.18	n.s.	-	0.77	1.00	0.91	0.90	n.s.	-	-
2b	Latent growth model with easy parcels, L1-4, linear. Slope var. @0	8.71	0.07	-	3.16	0**	-	0.75	0.89	0.91	0.93	-	-	-
3b	Latent growth model with easy parcels, L1-4, quadratic. Slope & quad. var. @0	8.70	n.s.	n.s.	3.16	0**	0**	0.75	0.89	0.91	0.93	-	-	-

** variance fixed at 0

When comparing slope parameters it is important to note the estimates for the linear and quadratic models represent change across one lag (e.g. from lesson 1 to lesson 2), while the slope estimate for the unconditional model represents change across all three lags (e.g. from lesson 1 to lesson 4). In the case of grade 6 the intercept was slightly higher than 9 (on a scale from 0 to 10) for all three models (Models 1b, 2b, and 3b), while the slope parameters indicated changes ranging from about half a score for the linear or unconditional models (Model 1b, .51; Model 2b, .48) to nearly three quarters of a score for the quadratic model (.72, Model 3b), across all four lessons. The intercept and slope covariance was significant in the case of the grade 6 unconditional model (standardised estimate, e.g. correlation = -.72), and marginally significant in the case of the linear and the quadratic models (standardised estimates, e.g. correlations, both -.65, $p<.10$). In the case of grade 10, the three models all estimated comparable intercepts (a bit below 9), while the slope parameters were either non-significant (unconditional and quadratic model) or very small (linear model). Overall, the estimated parameters were comparable across models, in the case of both groups.

For figures displaying growth models I have chosen the term "Level" to represent intercept, "Slope" to represent the slope factor for linear and quadratic models, and "Shape" to represent the slope factor for unconditional models. Figures 37 and 38 display the parameter estimates for the best-fitting grade 6 and grade 10 models, respectively. These models indicate that the younger students' self-efficacy grew slightly (slope = .16 between each lesson) in a linear fashion across the four lessons, while the grade 10 students' self-efficacy for easy tasks did not change significantly. There were significant variances in intercept (e.g. mean self-efficacy at the beginning of lesson 1) for both student groups, and marginally significant variances in slope in the case of the grade 6 sample. Furthermore, the covariance between self-efficacy intercept and slope for grade 6 was marginally significant (-.25), meaning there was a

tendency that in the case of grade 6 students, higher mean scores at lesson 1 was associated with lower estimates of slope, e.g. less steep changes over time.

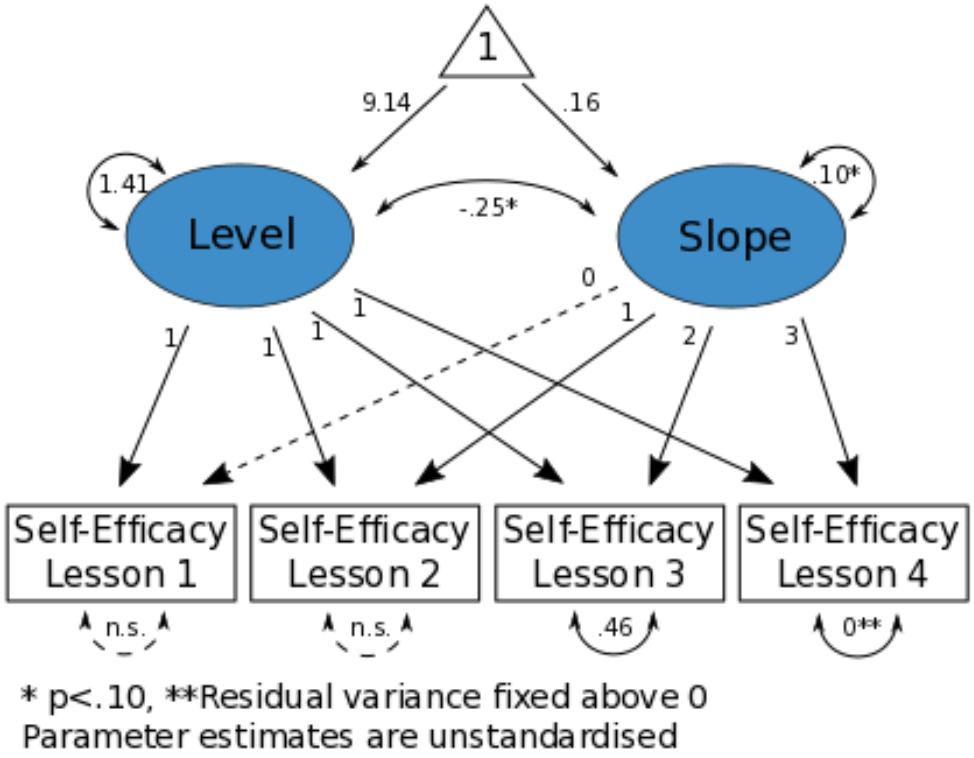


Figure 37: Linear growth model Self-Efficacy easy tasks, grade 6

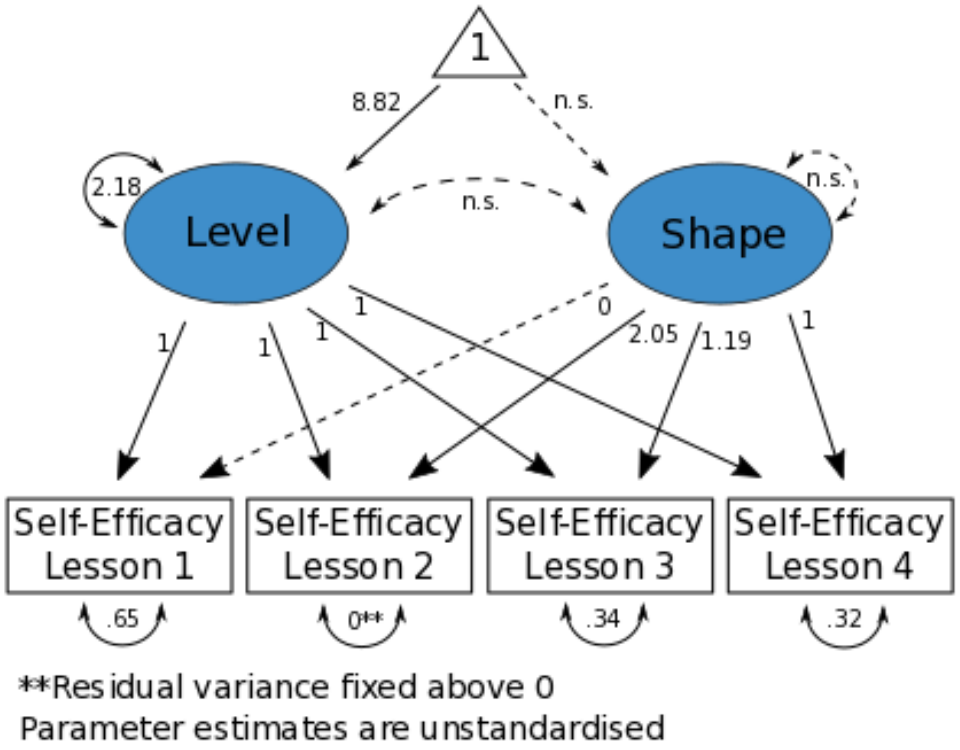


Figure 38: Unconditional growth model Self-Efficacy easy tasks, grade 10

Figure 39 displays the estimated means in students' self-efficacy for easy tasks, across the four lessons. The grade 6 growth-slope is estimated by the linear growth model, while the grade 10 slope is estimated on the basis of the unconditional growth model.

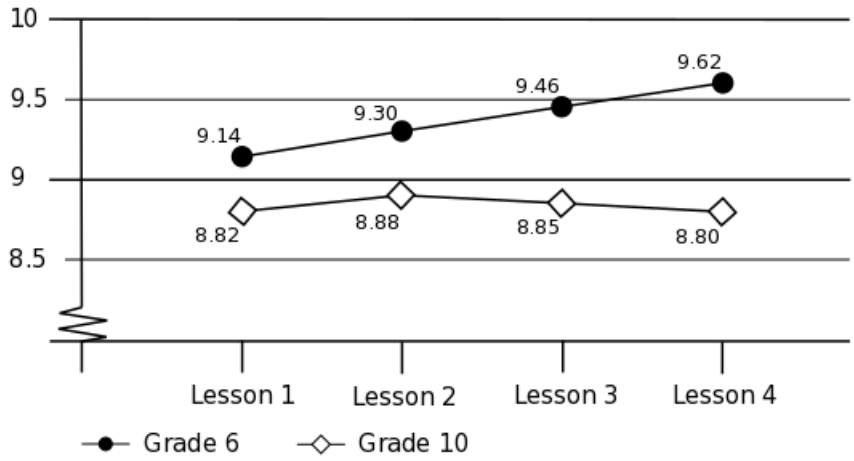


Figure 39: Estimated growth curves, Self-Efficacy easy tasks

Self-efficacy medium difficulty tasks

Tables 62 and 64 display model fit indices for the growth models for medium difficulty tasks. In the case of grade 6, the quadratic model fit best (Model 3b), while the unconditional model fit best for the grade 10 sample (Model 1). Considering the parameter estimates in Tables 63 and 65, the estimated intercept was again comparable across all three models, in the case of both samples (grade 6: just over 8, grade 10: about 7.50). In the case of grade 6, the slope parameters indicated a growth ranging from just under one score in self-efficacy in the case of the unconditional and the linear models (Models 1 and 2), to about one and three quarters score in the case of the quadratic model (Model 3b), across the four time points. In the case of grade 10, the unconditional and the linear model (Models 1 and 2b) estimated comparable growth parameters across the four time points (.30 and .43, respectively), while the quadratic model indicated no significant change. Similar to the easy level growth models, the intercept and slope covariance was non-significant in relation to the grade 10 models, while in the case of all the grade 6 models the intercept and slope covariance was significant (the standardised estimates, e.g. correlations, ranged from -.71 to -.80).

Table 62: Fit indices growth models medium difficulty tasks, grade 6

		Fit indices						*Model comparison*		
	Model	χ^2	*p*	RMSEA	CFI	TLI	SRMR	ΔCFI	$\Delta\chi^2_{(\Delta df)}$	*p*
1	Latent growth model with medium difficulty parcels, L1-4, unconditional	$\chi^2_{(3)}$= 5.70	*p*=.127	0.11	0.98	0.96	0.11			
2	Latent growth model with medium difficulty parcels, L1-4, linear	$\chi^2_{(5)}$= 8.11	*p*=.150	0.09	0.98	0.98	0.10	–0.003	$\Delta\chi^2_{(2)}$= 2.41	*p>.05*
3	Latent growth model with medium difficulty parcels, L1-4, quadratic**	$\chi^2_{(1)}$= .02	*p*=.889	0.00	1.00	1.04	0.00			
3b	**Latent growth model with medium difficulty parcels, L1-4, quadratic. Quad. var. @0**	**$\chi^2_{(4)}$= 2.94**	***p*=.567**	**0.00**	**1.00**	**1.01**	**0.10**	0.018	$\Delta\chi^2_{(1)}$= -2.76	*p>.05*

**Negative residual variance of quadratic factor

Table 63: Parameter estimates medium difficulty task models, grade 6

Unstandardized parameter estimates and model-explained variances

	Model	Fixed effects			Random effects			Variances explained				Covariances		
		Intercept	Slope	Quad.	Intercept	Slope	Quad.	L1	L2	L3	L4	I with S	I with Q	S with Q
1	Latent growth model with medium difficulty parcels, L1-4, unconditional	8.08	0.96	-	2.91	n.s.	-	0.92	0.90	0.72	0.82	-1.08	-	-
2	Latent growth model with medium difficulty parcels, L1-4, linear	8.19	0.31	-	2.72	n.s.	-	0.84	0.93	0.72	0.77	-0.31	-	-
3b	**Latent growth model with medium difficulty parcels, L1-4, quadratic. Quad. var. @0**	8.08	0.59	-0.09	2.75	.07*	0**	0.86	0.93	0.72	0.81	-0.33	-	-

*p<.10, ** variance fixed at 0

Table 64: Fit indices growth models medium difficulty tasks, grade 10

		Fit indices						*Model comparison*		
	Model	χ^2	*p*	RMSEA	CFI	TLI	SRMR	ΔCFI	$\Delta\chi^2_{(\Delta df)}$	*p*
1	**Latent growth model with medium difficulty parcels, L1-4, unconditional**	**$\chi^2_{(3)}$= 2.92**	***p*=.404**	**0.00**	**1.00**	**1.00**	**0.05**			
2	Latent growth model with medium difficulty parcels, L1-4, linear**	$\chi^2_{(5)}$= 10.49	*p*=.063	0.11	0.98	0.97	0.10			
2b	Latent growth model with medium difficulty parcels, L1-4, linear. Slope var. @0	$\chi^2_{(7)}$= 12.75	*p*=.078	0.09	0.98	0.98	0.14	–0.024	$\Delta\chi^2_{(4)}$= 9.83	*p<.05*
3	Latent growth model with medium difficulty parcels, L1-4, quadratic**	$\chi^2_{(1)}$= 3.40	*p*=.065	0.16	0.99	0.94	0.03			
3b	Latent growth model with medium difficulty parcels, L1-4, quadratic. Slope & quad. var. @0	$\chi^2_{(6)}$= 12.56	*p*=.051	0.11	0.97	0.97	0.14	–0.027	$\Delta\chi^2_{(3)}$= 9.64	*p<.05*

**Negative residual variance on one factor

Table 65: Parameter estimates medium difficulty task models, grade 10

Unstandardized parameter estimates and model-explained variances

	Model	Fixed effects			Random effects			Variances explained				Covariances		
		Intercept	Slope	Quad.	Intercept	Slope	Quad.	L1	L2	L3	L4	I with S	I with Q	S with Q
1	**Latent growth model with medium difficulty parcels, L1-4, unconditional**	7.36	0.43	-	3.91	n.s.	-	0.75	0.94	0.92	0.93	n.s.	-	-
2b	Latent growth model with medium difficulty parcels, L1-4, linear. Slope var. @0	7.54	0.10	-	5.28	0**	-	0.79	0.92	0.92	0.93	-	-	-
3b	Latent growth model with medium difficulty parcels, L1-4, quadratic. Slope & quad. var. @0	7.48	n.s	n.s	5.27	0**	0**	0.79	0.93	0.92	0.93	-	-	-

** variance fixed at 0

Figures 40 and 41 display the parameter estimates for the best-fitting growth models for each of the samples, in regard to the medium difficulty self-efficacy construct. The parameter estimates indicate that for both samples, students' self-efficacy for medium difficulty tasks increased significantly across the four lessons. In the case of grade 6 the estimated slope was .59 from one lesson to the next, while the grade 10 slope was .43 across all four time points. Furthermore, there was a significant quadratic, e.g. non-linear, term to the growth of the grade 6 students' self-efficacy scores, indicating a slightly flattened growth curve over time (-.09). The estimated factor loadings on the shape factor for the grade 10 sample indicate the older students' self-efficacy grew in a non-linear fashion, too, increasing from lessons 1 to 2 and 3 to 4, while decreasing from lesson 2 to 3. In the case of both samples there were significant differences in students' mean self-efficacy scores (intercept variance). For the grade 6 sample there were marginally significant differences in the slope of students' self-efficacy (slope variance). Furthermore, there was a significant relationship between grade 6 students' mean self-efficacy at the beginning of lesson 1, and the steepness of the growth slope, where higher mean self-efficacy was related to a less steep slope (intercept with slope covariance = -.33). For the older sample, this relationship was not significant.

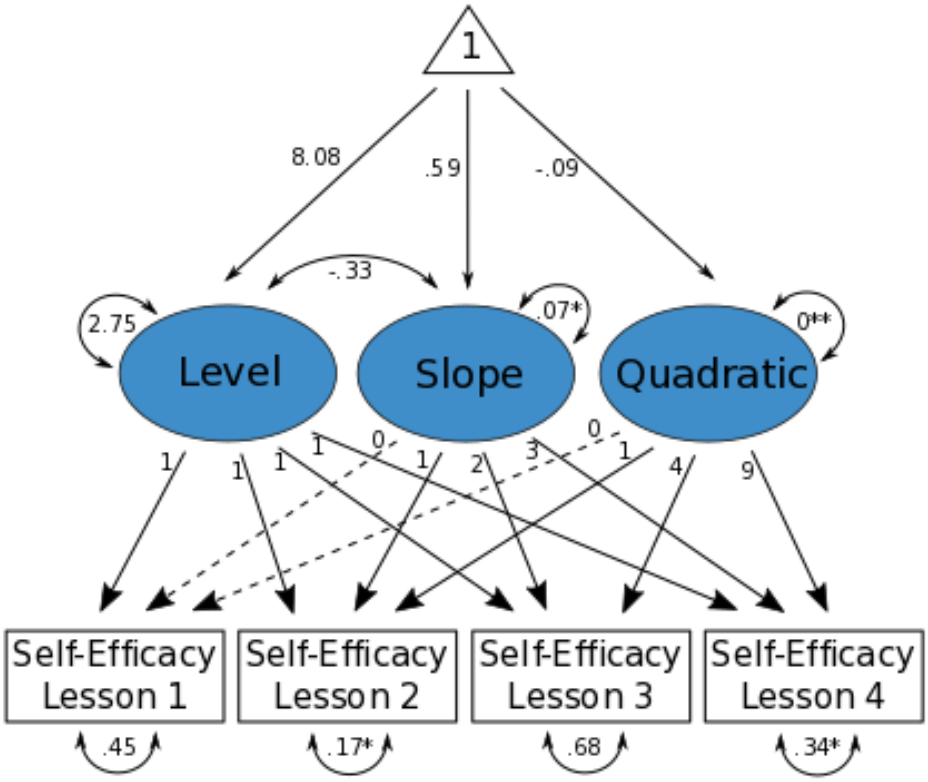


Figure 40: Quadratic growth model medium difficulty tasks, grade 6

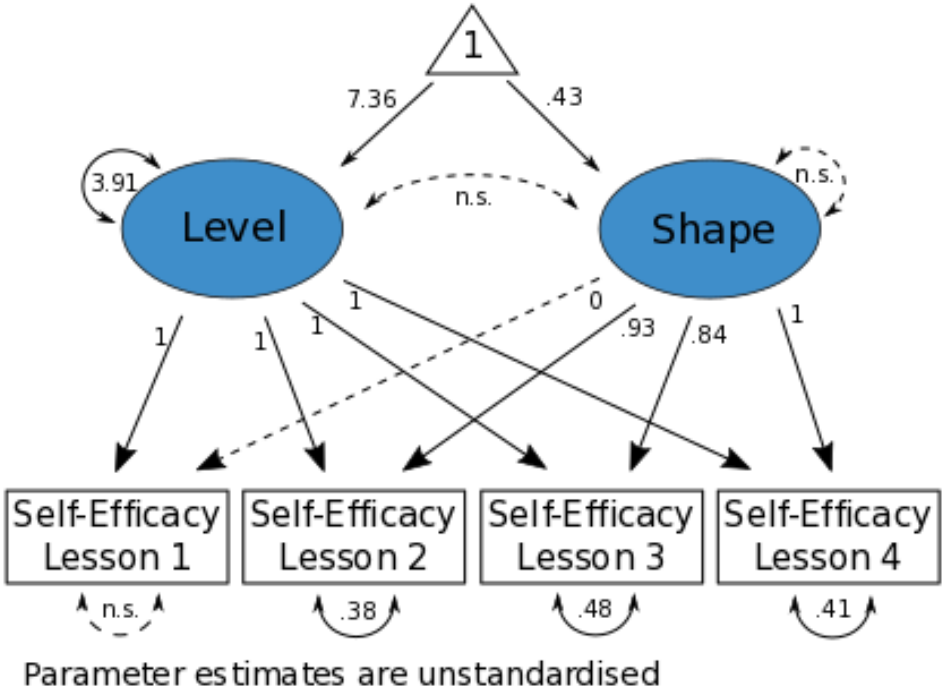


Figure 41: Unconditional growth model medium difficulty tasks, grade 10

Figure 42 displays the estimated mean self-efficacy scores for medium difficulty tasks, with grade 6 and grade 10 on separate lines.

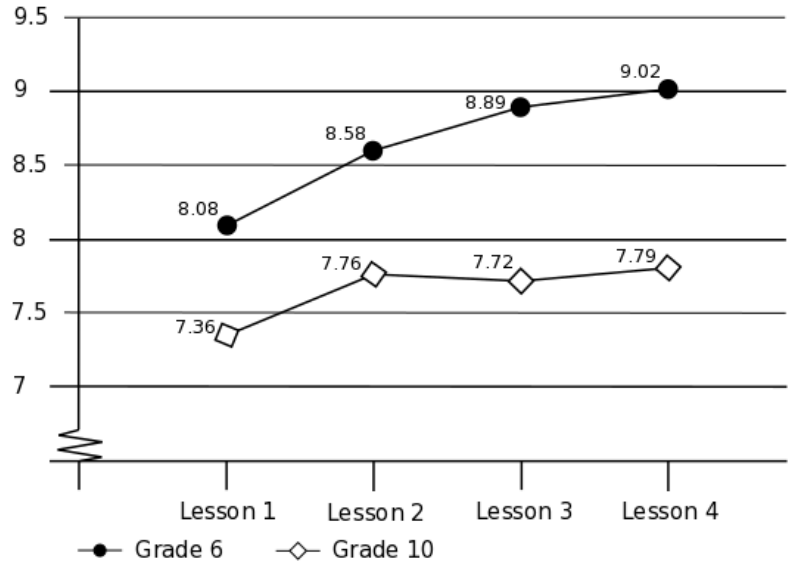


Figure 42: Estimated growth curves, Self-Efficacy medium difficulty tasks

Self-efficacy hard tasks

Tables 66 and 68 display the fit indices for the nested growth models for hard tasks for grade 6 and grade 10, respectively. In the case of self-efficacy for hard tasks, the quadratic model (Model 3b) fit best for grade 6, while the unconditional model (Model 1) fit best in regard to the grade 10 sample. Considering the χ^2 difference only, the linear model (Model 2) fit marginally better in regard to the grade 10 sample – however considering CFI as well as wider fit indices, the unconditional model fit clearly better. I present the parameter estimates derived from all three alternative models in Tables 67 and 69. In the case of both grades, the estimated intercepts were comparable across the three models (between 5.65 and 5.83 in the case of grade 6, and about/just below 5 in the case of grade 10). In the case of grade 6, the slope parameters ranged from a change in self-efficacy of about one and a half score for the unconditional and linear models (Models 1 and 2) to about three scores in the case of the quadratic model (Model 3b), across the four lessons. In the case of grade 10, the slope ranged from just below one score for the unconditonal and linear models (Models 1 and 2) to about

one and three quarters score in the case of the quadratic model (Model 3), across all time points. Intercept and slope covariances were not significant in any of the models, for either of the samples.

Table 66: Fit indices growth models hard tasks, grade 6

		Fit indices						*Model comparison*		
	Model	χ^2	*p*	RMSEA	CFI	TLI	SRMR	ΔCFI	$\Delta\chi^2_{(\Delta df)}$	*p*
1	Latent growth model with hard parcels, L1-4, unconditional	$\chi^2_{(3)}$= 9.68	*p*=.022	0.17	0.95	0.90	0.07			
2	Latent growth model with hard parcels, L1-4, linear	$\chi^2_{(5)}$= 14.54	*p*=.013	0.15	0.93	0.91	0.06	–0.022	$\Delta\chi^2_{(2)}$= 4.86	*p>.05*
3	Latent growth model with hard parcels, L1-4, quadratic*	$\chi^2_{(1)}$= .64	*p*=.422	0.00	1.00	1.02	0.01			
3b	**Latent growth model with hard parcels, L1-4, quadratic. Quad. var. @0**	**$\chi^2_{(4)}$= 7.35**	***p*=.119**	**0.10**	**0.97**	**0.96**	**0.07**	0.025	$\Delta\chi^2_{(1)}$= -2.33	*p>.05*

*Negative residual variance of one item

Table 67: Parameter estimates hard task models, grade 6

Unstandardized parameter estimates and model-explained variances

	Model	Fixed effects			Random effects			Variances explained				Covariances		
		Intercept	Slope	Quad.	Intercept	Slope	Quad.	L1	L2	L3	L4	I with S	I with Q	S with Q
1	Latent growth model with hard parcels, L1-4, unconditional	5.66	1.47	-	5.06	n.s.	-	0.91	0.81	0.90	0.85	n.s.	-	-
2	Latent growth model with hard parcels, L1-4, linear	5.83	0.52	-	4.63	0.10*	-	0.79	0.82	0.84	0.87	n.s.	-	-
3b	**Latent growth model with hard parcels, L1-4, quadratic. Quad. var. @0**	5.65	1.02	-0.17	4.71	0.13	0**	0.82	0.82	0.85	0.90	n.s.	-	-

*p<.10, ** variance fixed at 0

Table 68: Fit indices growth models hard tasks, grade 10

		Fit indices						*Model comparison*		
	Model	χ^2	*p*	RMSEA	CFI	TLI	SRMR	ΔCFI	$\Delta\chi^2_{(\Delta df)}$	*p*
1	**Latent growth model with hard parcels, L1-4, unconditional**	**$\chi^2_{(3)}$= 3.55**	***p*=.314**	**0.04**	**1.00**	**0.99**	**0.03**			
2	Latent growth model with hard parcels, L1-4, linear	$\chi^2_{(5)}$= 7.83	*p*=.166	0.08	0.98	0.98	0.04	–0.013	$\Delta\chi^2_{(2)}$= 4.28	*p>.05*
3	Latent growth model with hard parcels, L1-4, quadratic	$\chi^2_{(1)}$= 1.12	*p*=.290	0.04	1.00	1.00	0.01	0.002	$\Delta\chi^2_{(2)}$= -2.43	*p>.05*

Table 69: Parameter estimates hard task models, grade 10

Unstandardized parameter estimates and model-explained variances

	Model	Fixed effects			Random effects			Variances explained				Covariances		
		Intercept	Slope	Quad.	Intercept	Slope	Quad.	L1	L2	L3	L4	I with S	I with Q	S with Q
1	**Latent growth model with hard parcels, L1-4, unconditional**	4.86	0.87	-	5.88	n.s.	-	0.85	0.88	0.85	0.85	n.s.	-	-
2	Latent growth model with hard parcels, L1-4, linear	5.02	0.28	-	5.92	n.s.	-	0.80	0.87	0.84	0.87	n.s.	-	-
3	Latent growth model with hard parcels, L1-4, quadratic	4.88	0.59	n.s.	6.10	n.s.	n.s.	0.89	0.89	0.87	0.85	n.s.	n.s.	n.s.

Figures 43 and 44 display the parameter estimates for the best-fitting growth models for students' self-efficacy for hard tasks. These estimates indicate students' self-efficacy for hard tasks grew significantly across the four lessons, for both samples (grade 6 slope 1.02 between two adjacent lessons, grade 10 slope .87 from the first to the last lesson). As was the case for the self-efficacy for medium difficulty tasks, this growth decreased over time in the case of the grade 6 sample (quadratic term -.17). Considering the factor loadings for the shape factor for the grade 10 sample, the growth in self-efficacy was steeper from the first to the second lesson (as compared with subsequent lessons) for the older students, too. For both samples, there were significant differences in students' mean self-efficacy at the beginning of the first lesson (intercept variance), and in the case of the grade 6 sample there were significant differences in students' slope, also (slope variance).

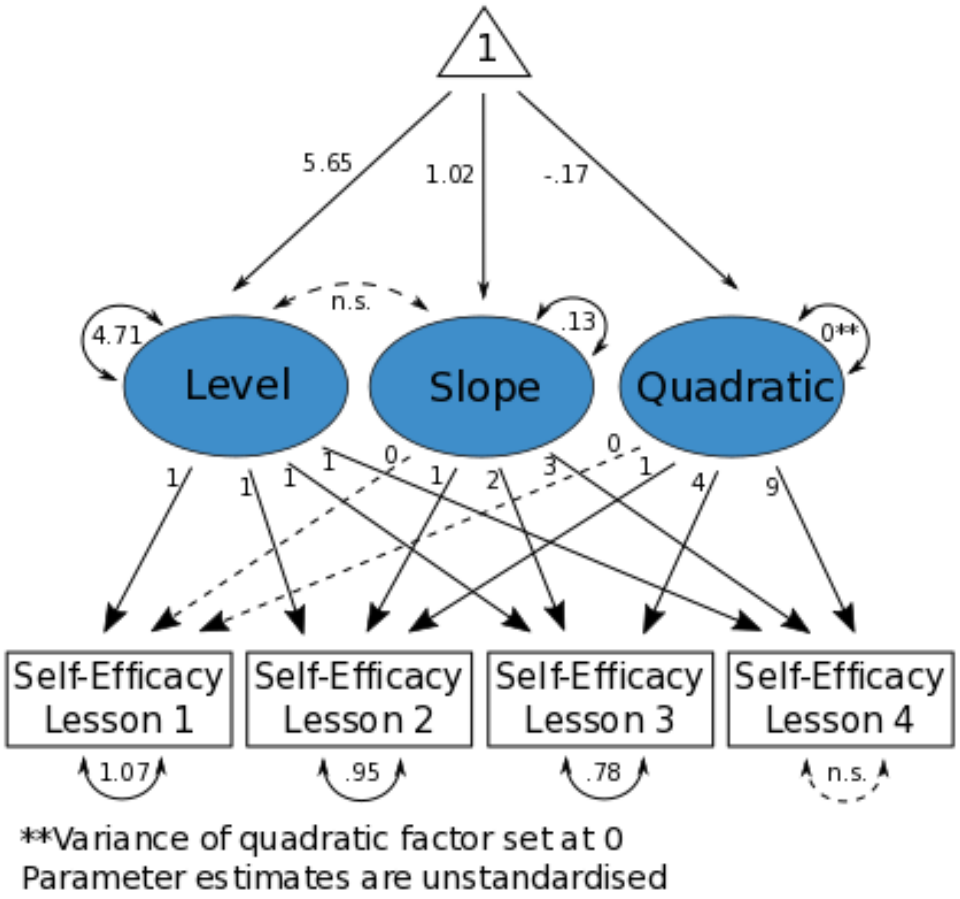


Figure 43: Quadratic growth model hard tasks, grade 6

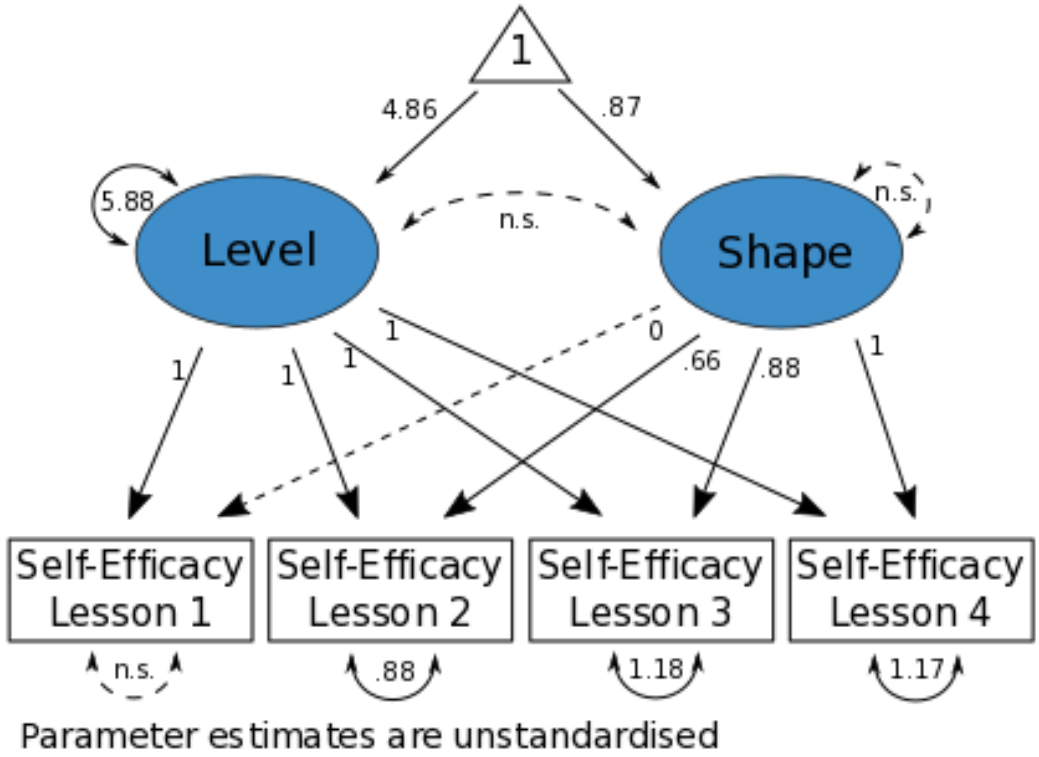


Figure 44: Unconditional growth model hard tasks, grade 10

Figure 45 displays the estimated mean scores of students' self-efficacy for hard tasks, at each of the lesson time points, based on the quadratic growth model.

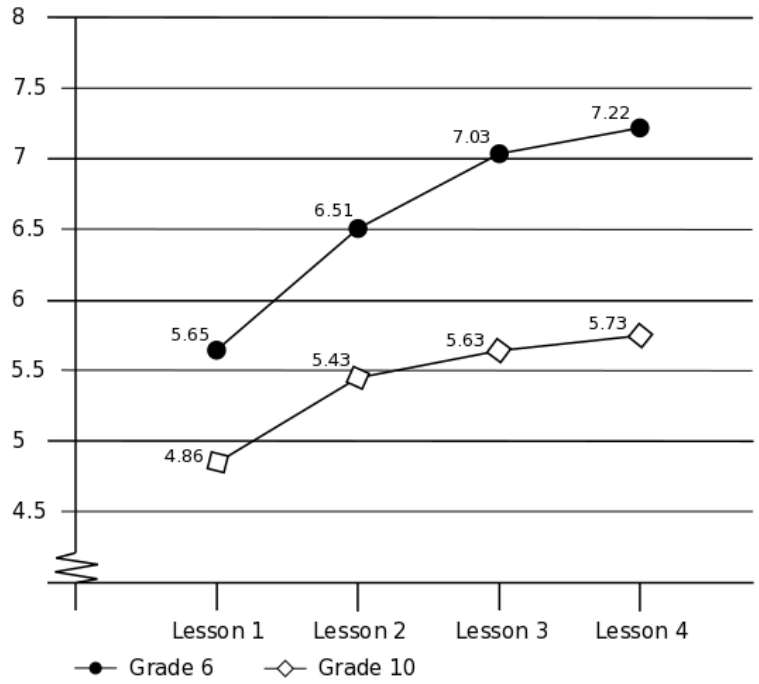


Figure 45: Estimated growth curves, Self-Efficacy hard tasks

Differences in slope and intercept between grade 6 and grade 10

In the case of all three self-efficacy constructs, grade 6 intercept and slope estimates were higher, in absolute terms, as compared to the grade 10 estimates. In order to investigate whether these differences were statistically significant, I specified group-as-moderator models, and imposed equality constraints on the slope and intercept factors. Given that different models were best-fitting for the two groups, I used the unconstrained model as a basis for group comparisons, as this allowed for separate estimations of the slope factor for each group. My aim in this instance was not to find the best-fitting model across the two groups, as I had already ascertained which model fit best for each of the groups separately. Rather, the aim was to investigate whether the intercept and slope of each group differed significantly. I thus found it reasonable to choose the model that would allow the best-fitting (but not necessarily most parsimonious) slope parameters, rather than looking for the best-fitting group model in terms of fit indices, where the data from one of the groups would be "forced" into a model that I knew did not provide the best fit for that group. Tables 70 to 72 present the fit indices from the unconditional group-as-moderator models with freely estimated slope factors, for easy, medium difficulty, and hard tasks. In each of the tables Model 1 represents the baseline model, against which the models with equality constraints are compared. Models 2 and 3 each include equality constraints to investigate whether slope and intercept, respectively, differed significantly between grade 6 and grade 10 students.

In the case of the easy tasks, imposing equality constraints across the grades did not lead to significant detriment in fit, as demonstrated in Table 70. This indicates that in the case of the easy self-efficacy construct, students' self-efficacy did not differ, in regard to mean level at lesson 1 (intercept) or growth across the four lessons (slope). In the case of the medium

difficulty as well as hard tasks, imposing equality constraints for the slope or the intercept did lead to significant detriment to model fit (Tables 71 and 72). This indicates both the slope and intercept differed significantly between the grade 6 and grade 10 students, in regard to medium difficulty and hard tasks. In other words, grade 6 students' self-efficacy was significantly stronger at the beginning of lesson 1 (intercept), and changed in a steeper fashion across the four lessons (slope), as compared to the self-efficacy of students in grade 10.

Table 70: Group models with equality constraints, easy tasks

Group-as-moderator models, Easy tasks	*Fit indices*						*Model comparison*			
Model	χ^2	*p*	RMSEA	CFI	TLI	SRMR	ΔCFI	$\Delta\chi^2_{(\Delta df)}$	*p*	*pass?*
1 Unconditional growth model with easy level parcels, L1-4, allowing different slope factors.	$\chi^2_{(8)}$= 17.21	*p*=.028	0.11	0.95	0.93	0.27				
2 Unconditional growth model with easy level parcels, L1-4, allowing different slope factors. Equality constraints on slope	$\chi^2_{(9)}$= 19.43	*p*=.022	0.12	0.95	0.93	0.28	-0.006	$\Delta\chi^2_{(1)}$= 2.22	*p>.05*	*yes*
3 Unconditional growth model with easy level parcels, L1-4, allowing different slope factors. Equality constraints on intercept	$\chi^2_{(9)}$= 19.92	*p*=.018	0.12	0.95	0.93	0.26	-0.008	$\Delta\chi^2_{(1)}$= 2.71	*p>.05*	*yes*

Table 71: Group models with equality constraints, medium difficulty tasks

Group-as-moderator models, Medium difficulty tasks	*Fit indices*						*Model comparison*			
Model	χ^2	*p*	RMSEA	CFI	TLI	SRMR	ΔCFI	$\Delta\chi^2_{(\Delta df)}$	*p*	*pass?*
1 Unconditional growth model with medium difficulty level parcels, L1-4, allowing different slope factors.	$\chi^2_{(6)}$= 8.10	*p*=.231	0.06	1.00	0.99	0.09				
2 Unconditional growth model with medium difficulty level parcels, L1-4, allowing different slope factors. Equality constraints on slope	$\chi^2_{(7)}$= 15.21	*p*=.033	0.12	0.98	0.97	0.10	-0.016	$\Delta\chi^2_{(1)}$= 7.11	*p<.05*	*no*
3 Unconditional growth model with medium difficulty level parcels, L1-4, allowing different slope factors. Equality constraints on intercept	$\chi^2_{(7)}$= 17.55	*p*=.014	0.13	0.97	0.96	0.13	-0.021	$\Delta\chi^2_{(1)}$= 9.45	*p<.05*	*no*

Table 72: Group models with equality constraints, hard tasks

Group-as-moderator models, Hard tasks		*Fit indices*						*Model comparison*			
	Model	χ^2	*p*	RMSEA	CFI	TLI	SRMR	ΔCFI	$\Delta\chi^2_{(\Delta df)}$	*p*	*pass?*
1	Unconditional growth model with hard level parcels, L1-4, allowing different slope factors.	$\chi^2_{(6)}$= 12.59	*p*=.050	0.11	0.98	0.96	0.05				
2	Unconditional growth model with hard level parcels, L1-4, allowing different slope factors. Equality constraints on slope	$\chi^2_{(7)}$= 17.64	*p*=.014	0.13	0.97	0.94	0.07	-0.013	$\Delta\chi^2_{(1)}$= 5.05	*p<.05*	*no*
3	Unconditional growth model with hard level parcels, L1-4, allowing different slope factors. Equality constraints on intercept	$\chi^2_{(7)}$= 17.85	*p*=.013	0.13	0.97	0.94	0.09	-0.013	$\Delta\chi^2_{(1)}$= 5.26	*p<.05*	*no*

4.4.4.3 Summary of results for RQ3

I specified variance component models to investigate the amount of variance at the within (e.g. individual) versus the between (e.g. class) level. These models indicated most of the variance was as the individual level, while the ICCs (between-level variance) ranged from 0.3% to 18.6%. Considering the three self-efficacy constructs, more of the variance was at the between level for the hard self-efficacy construct, followed by the medium and easy self-efficacy constructs. Furthermore, across the three self-efficacy constructs, the lowest ICC (e.g. between level variance) was in relation to the second lesson. I inspected the mean self-efficacy in each of the classes, and found that students in one grade 10 class (class 11) reported significantly lower mean self-efficacy scores, as compared to the other classes, across several of the time points /levels of difficulty. This might indicate stable differences between this class and the other classes, in terms of teacher, school, or home influences. Interestingly, class number 10, which I identified in section 4.3.2.1 as above average in regard to mean scores on national tests during phase 1, did not report higher self-efficacy as compared to the other classes in phase 2.

In order to investigate mean changes in students' self-efficacy across lessons in mathematics, I specified a series of nested latent growth models. Given theoretical assumptions that changes in self-efficacy may differ as a function of student age (see section 4.2.1), I specified separate models for grade 6 and grade 10 students. Indeed, different growth shapes provided the best model fit for the two samples, considering a priori determined fit indices. In the case of the grade 6 students, a linear (self-efficacy easy tasks) and two quadratic models (self-efficacy medium difficulty and hard tasks) fit the data best, while in the case of the grade 10 students the unconditional model fit best across the easy, medium difficulty, and hard tasks. The shape (=growth) factor-loadings for the grade 10 unconditional models indicated a

steeper change from lesson 1 to lesson 2, than between the subsequent lessons, for all three self-efficacy constructs. Similarly, the quadratic term for the grade 6 medium difficulty and hard models indicated a flattening growth rate. Accordingly, a non-linear growth shape of self-efficacy, characterised by a rapid initial growth, followed by a flattening, seemed to best describe the growth of students' self-efficacy for both samples, across all three self-efficacy constructs (with the exception of the linear model in the case of grade 6 students' self-efficacy for easy tasks).

The intercept and slope variances provide information about the degree to which the mean level and growth of students' self-efficacy differed within each group. The variance of intercept was significant for both grades, across all task difficulties, demonstrating there was significant variation between students' mean level self-efficacy at the beginning of the first lesson. Furthermore, the estimate of intercept variance was higher (in an absolute sense) in relation to the grade 10 students as compared with grade 6, and increased with task difficulty in the case of both groups. A similar tendency was indicated by the slope variance for grade 6 students, where there was more variation in growth shape in regard to hard tasks, as compared with medium difficulty and easy tasks (the slope variance was marginally significant for easy and medium difficulty tasks). The variance of slope (shape) was non-significant in regard to grade 10 students, for all task difficulties. The relationship between intercept and slope (covariance) was significant in the case of grade 6 students' self-efficacy for easy and medium difficulty tasks, respectively, meaning grade 6 students with stronger self-efficacy at the beginning of the first lesson were likely to experience less growth in the construct across the lessons. This relationship was not significant in the case of grade 6 students' self-efficacy for hard tasks, or for grade 10 students' self-efficacy for tasks of any difficulty (as there was no significant variation in the slope of grade 10 students' self-efficacy).

Consistently, I found higher parameter estimates for both intercept (mean scores at the beginning of lesson 1) and slope (growth rate between lesson 1 and lesson 4) for the grade 6 models, as compared with the grade 10 models. For each of the latent self-efficacy constructs these differences were tested for statistical significance, through the imposition of equality constraints on unconditional group-as-moderator models. In the case of students' self-efficacy for medium and hard difficulty tasks, the differences were statistically significant in regard to both slope and intercept. This means in the case of medium difficulty and hard tasks, grade 6 students had higher self-efficacy at the beginning of lesson 1, and also steeper growth curves across the four lessons, as compared with students in grade 10. Finally, for both student samples self-efficacy intercept (level) was progressively weaker, and self-efficacy slope (growth) was progressively steeper with increasing task difficulty. This is demonstrated visually in Figure 46. These differences could not be tested statistically, as the three latent self-efficacy constructs were investigated in separate models.

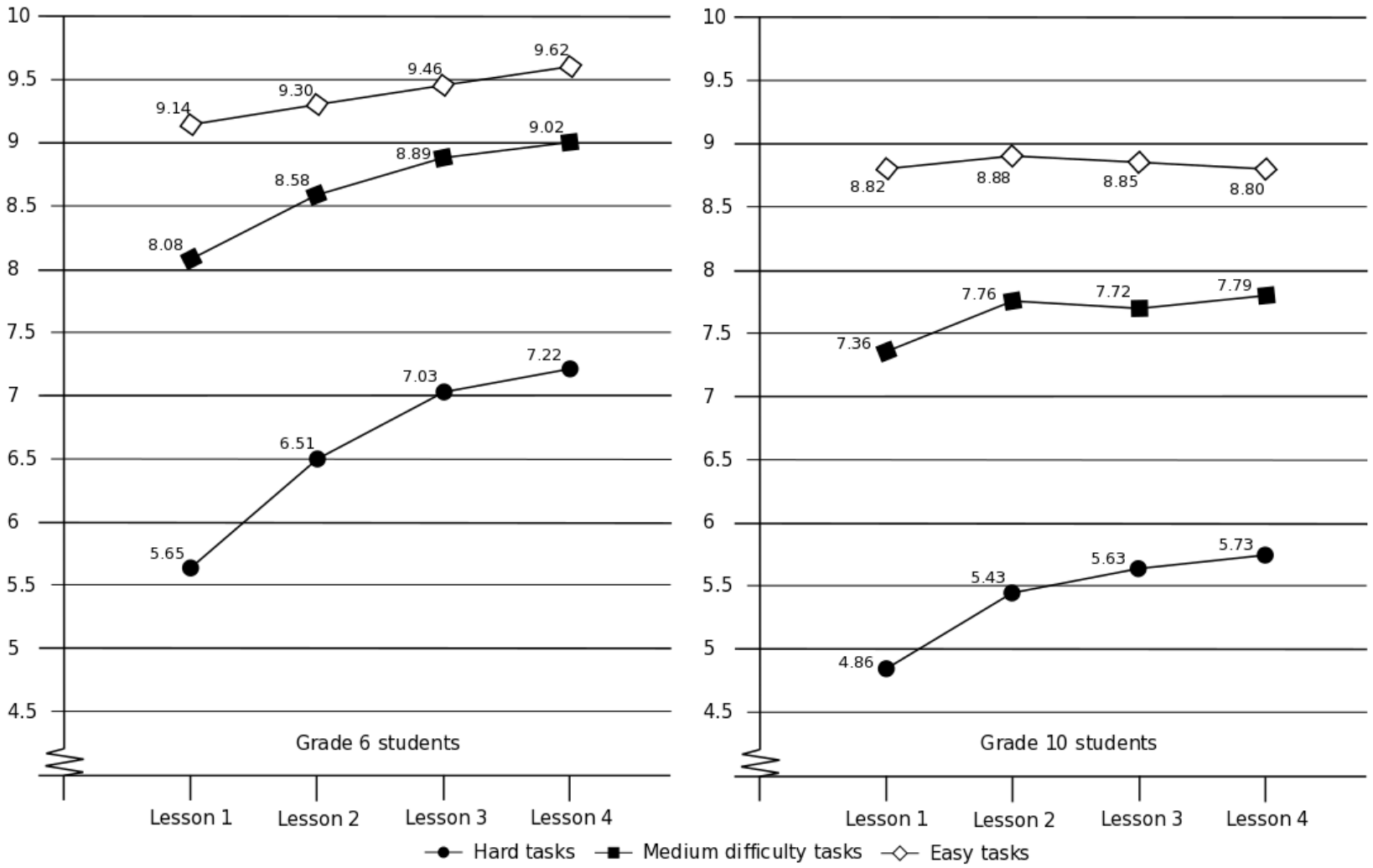


Figure 46: Estimated growth curves for Self-Efficacy, grade 6 and grade 10

4.5 Discussion

In the present study, stability and change of students' self-efficacy have been investigated using descriptive statistics, autoregressive path models, and latent growth models for change. These analytical approaches have provided complimentary findings regarding changes in students' self-efficacy in mathematics during a sequence of four mathematics lessons when a new topic was introduced. Knowledge about changes in self-efficacy across short measurement lags can further our theoretical understanding of the process of self-efficacy change, as well as provide important implications for classroom practice. I will now discuss my findings, starting with the descriptive analyses before discussing my key findings from each research question.

4.5.1 Reliable changes in self-efficacy

The descriptive statistics provided from the reliable change scores were quite informative, and the indications obtained from these statistics proved consistent with the findings from the more advanced analytical approaches. The reliable change scores indicated that students' self-efficacy does change. Across all lessons, between 15 – 30 % of students in my study reported one or more reliable change, lower than the 60 % reported by Bernacki et al. (2015). This difference may be related to the different measures of self-efficacy. I used a latent construct, and operationalised self-efficacy at a more general (*lesson*) level, in comparison to the single, and more specific, self-efficacy *task* item used by Bernacki et al. The distance between the measurement occasions may also be important in this respect – in the study by Bernacki et al. (2015) self-efficacy was measured four times with short lags (approximately 5 minutes apart), while in my study the lag between adjacent measurement occasions ranged from 45 minutes to 5 days. It will be important to unpack these potential influences in future studies, using

advanced statistical methods, to investigate the relative contribution to change by the degree of specificity of the self-efficacy construct, and the lag between measurement occasions.

For grade 6, the reliable change scores in my study indicated that changes in self-efficacy predominantly occurred early on after the teacher introduced a new topic. This highlights the importance of the design of my study, where I followed students across a sequence of lessons as they engaged with a new topic in mathematics. The reliable change scores also indicated differences according to levels of difficulty, where a higher percentage of grade 6 students reported "reliable changes" in association with hard level tasks, as compared with easy and medium difficulty tasks. The highest percentage of "reliable changes" for grade 10 students' self-efficacy was in relation to tasks of medium difficulty. These findings indicate changes in self-efficacy are linked with both the novelty and difficulty of tasks. While these findings illustrate the value of presenting "simple" descriptive statistics, they tell us little about overall trends in the data, which I will discuss in view of the findings from my research questions.

4.5.2 A multidimensional measure of self-efficacy for lessons in mathematics

The findings from the first research question in this study replicated my previous findings in study 1, with a multidimensional structure of self-efficacy. Compared with plausible alternatives, and consistent with Study 1, a factor structure including three latent level constructs of self-efficacy provided the best fit to my data, and my three self-efficacy level constructs were supported as strong or partially strong factorially invariant. This strengthens my confidence in the findings from study 1 and my argument that, in addition to facets of mathematics, students consider the level of difficulty of mathematical tasks when forming their self-efficacy.

To my knowledge, no previous measures including both facet-specificity and levels of difficulty for students' self-efficacy regarding lessons in mathematics exist, and my proposed measure has theoretical as well as practical implications. While Bandura asserts that levels of difficulty are integral to self-efficacy (e.g. *regardless of the level of generality, in no case are the efficacy items dissociated from context and level of task demands*) (1997, p. 50), in most research of which I am aware levels of difficulty have not been explicitly considered (see study 1, section 2.2.3 for notable exceptions). If self-efficacy is treated as de-contextualised in regard to varying levels of difficulty, this could imply that task demands are seen as external influences on relatively stable (in regard to difficulty) constructs of self-efficacy. In contrast to such an interpretation and in line with the assertion by Bandura, I propose that levels of perceived task difficulty are part of, rather than separate to, students' self-efficacy. In line with this theoretical stance, in my measure I proposed a novel operationalisation of self-efficacy, where the theoretically proposed levels of difficulty are explicitly included. For researchers, including levels of difficulty in investigations about self-efficacy can provide new insights into how self-efficacy develops across task engagement, as well as whether / how levels of difficulty interact with the relationship between self-efficacy and other related constructs, such as sources of self-efficacy. As such, it can be an important step in expanding our knowledge of factors that moderate the theoretically proposed process of self-efficacy change.

From an instructional point of view, my findings imply that levels of perceived difficulty matter to students when they form their self-efficacy. As such, it might be beneficial for students to have support from the teacher in developing their skills in assessing task difficulty, and in selecting adequately challenging tasks. The issue of what level of task difficulty is "adequate", in regard to facilitating self-efficacy change, was addressed in research questions

2 and 3, in the current study, and in study 4, concerning the relationship between students' self-efficacy and their mastery experiences.

4.5.3 High rank-order stability of students' self-efficacy

In research question 2, autoregressive path models provided information about the stability of the rank-order of students, as determined by their self-efficacy scores for each of the lessons in mathematics. My analyses gave rise to the following key findings:

1) the simplex model of change provided a good fit to my data;
2) the stability of the rank-order of students' self-efficacy was high when learning a new topic;
3) the stability of the rank-order did not change significantly across the lessons; and
4) the stability of the rank-order appeared similar in relation to different levels of perceived difficulty.

I will now discuss each of these key findings.

For key finding one, the simplex model provided a good fit to the data for all three levels of perceived difficulty. The simplex process indicates that students' self-efficacy at any given time point was more strongly related to the immediately proceeding time point, than to their initial self-efficacy. As pointed out by Shea and Howell (2000) this is in support of Bandura's theoretical proposition of self-efficacy as a dynamic construct, which changes over time according to the most recent self-efficacy changing experiences. This implies teachers should be able to support changes in students' self-efficacy through facilitating "self-efficacy changing experiences", such as mastery experiences during lessons of mathematics. I will return to the relationship between students' mastery experiences and their self-efficacy in study 4.

Considering the second and third key findings, the simplex models demonstrated that the stability was high in regard to all three self-efficacy level constructs (standardised estimates ranging from .85 to .98,), and that the rank-order stability did not change across the lessons. For each of the self-efficacy constructs the unstandardised path estimates were .94 (easy), .96 (medium difficulty) and .92 (hard). The differences between the stability estimates for each of the three latent self-efficacy constructs were relatively small, and were not investigated statistically.

The stability of students' self-efficacy was higher in my study than in previous self-efficacy studies of which I am aware, where the regression paths in autoregressive models have typically ranged from .30 to .40 (see section 4.2.1). This might be related to differing time frames. Previous studies have generally measured self-efficacy across months or years, while I followed students' self-efficacy for about a week. The rank-order of students' self-efficacy might change as a function of accumulated mastery experiences over time. Furthermore, as mentioned in section 4.2, cognitive maturation might be related to changes in self-appraisals, and this might result in lower self-efficacy stability estimates across longer time frames. Also, previous studies have provided evidence of self-efficacy change linked with exam times (e.g. Bong, 2005) or periods of transition (e.g. school changes, Friedel et al., 2010). Although in my study students were introduced to a new topic, their lessons were otherwise "regular" lessons in mathematics, and there were no immediately upcoming exams.

While the differences with previous studies might in part be explained by methodological differences, the high stability of self-efficacy in my study was somewhat unexpected, particularly considering students were introduced to a new mathematical topic. Theoretically, self-efficacy changes are linked with influential events such as experiences of mastery

(Bandura, 1997), and theory predicts self-efficacy is most likely to change when students learn new things (Usher & Pajares, 2008).

Furthermore, in terms of key finding four, the rank-order stability of students' self-efficacy was high for all levels of task difficulty. It was somewhat unexpected that there were such small differences between the rank-order stability of self-efficacy for different levels of tasks, considering that task difficulty is related to whether students are able to succeed on tasks or not, and thereby related to students' mastery experiences. This finding also stands in contrast to my findings in research question 1, where I argued that students consider the level of perceived difficulty when forming their self-efficacy.

It seems that the experiences available to students across the sequences of lessons investigated in my study, did not provide the kinds of "influential events" sufficient to disrupt the rank-order of students' self-efficacy to a large degree, for any level of task difficulty. As discussed in section 4.2 experiences with mathematics do not influence students' self-efficacy directly, but rather are cognitively appraised. It is possible that high stability of students' *cognitive appraisals* of their experiences can explain how the presumed availability of performance experiences from engaging with new tasks did not lead to a higher disruption of the rank-order. Butz and Usher (2015, p. 60) argued that students who have come to doubt their academic capabilities might be prone to look for "confirming evidence of their own inefficacy" – I would argue a similar tendency of looking for confirming evidence is likely in the case of students who see themselves as highly capable in mathematics.

If the stability of students' cognitive appraisals of their experiences causes higher stability to their self-efficacy than what might otherwise have followed from their "actual" performances,

then this has implications for how teachers work with performance feedback. I would argue that beyond supporting students' skill development in order to succeed, teacher support on how to interpret and make sense of performance experiences might be key for students to benefit from positive ("successful") performances. This is perhaps especially pertinent in the case of students who have come to doubt their mathematical abilities. In study 4, research question 3, I investigate the relationship between students' self-efficacy and their ability conceptions, which are related to students' cognitive appraisals of performance experiences.

Past studies have argued that due to a lack of previous first-hand experiences, students are particularly likely to rely on social comparison information in the case of new tasks (e.g. Pietsch, Walker, & Chapman, 2003, p. 597), which might also have worked to strengthen the rank-order stability in my findings. However, on the basis of this I would have expected decreasing stability across the lesson sequence, as students gained first-hand experiences with the tasks, which was not evident in my sample. Finally, Talsma et al. (2018) reported higher stability of self-efficacy for adults (.69), than for children (.36). In accordance with this, I would expect also higher stability of self-efficacy for the older, as compared with the younger, students, which was not tested in the current research question.

Limitations

The stability estimates were in relation to one sample including two groups of students (grade 6 and grade 10), and might as such hide group-level (as well as individual-level) variation, which is a limitation of my findings. I chose to include both groups in the autoregressive models, while including grade level as a covariate. In line with previous research (Bong, 2009), grade predicted self-efficacy at lesson 1 negatively, meaning older students were more likely to hold weaker self-efficacy. Previous research has reported higher accuracy of self-

efficacy for older students (Bong, 2009; Multon et al., 1991), and it is possible self-efficacy is also more stable with age, as per the findings by Talsma et al. (2018). Growing stability with increasing age has been reported in relation to self-concept (Davis-Kean et al., 2008), competence beliefs (Wigfield et al., 1997), and intrinsic motivation in mathematics (Gottfried et al., 2001). Gottfried et al. (2001) argued that rank-order stability might increase as a function of age across motivational domains. To my knowledge this has not been tested in relation to self-efficacy, and should be investigated in future studies. My findings from research question 2 were complimented by my findings from research question 3, where I investigated changes in mean-level self-efficacy separately for students in grade 6 and grade 10.

4.5.4 Mean-level changes in self-efficacy

To investigate changes in mean self-efficacy across a sequence of mathematics lessons I used latent growth modelling. These investigations demonstrated that:

1) students' self-efficacy changed significantly across a sequence of lessons;
2) the growth was characterised by an initially steeper incline, followed by a flattening trend;
3) there were significant differences between grade 6 and grade 10 students' intercept and growth; and
4) the slope of students' self-efficacy was steeper and the intercept lower in association with higher task difficulty.

I will discuss each of these key findings next.

In relation to key findings 1 and 2, students' self-efficacy grew in a non-linear fashion, with a steeper initial change in self-efficacy followed by a flattening growth curve. I interpret this shape of growth in relation to the timing of the study, where the first time point was measured

at the start of a lesson when students were introduced to a new mathematical topic. As such, my findings give empirical support to the assertion that self-efficacy is most likely to change when students work on new tasks (e.g. Usher & Pajares, 2008). On a group level, the steepest growth in students' self-efficacy occurred between the beginning of lesson 1 and the beginning of lesson 2, when students had engaged with the new topic for a single mathematics lesson. It is likely that the initial stages of learning provide opportunities for novel mastery experiences, and as such are opportune moments to affect positive changes in students' self-efficacy. If the "novelty" of the mastery experiences is key to affecting self-efficacy change, this has implications for teacher - student feedback. If teachers highlight differences and improvement from previous performances when giving feedback, this might increase the salience of students' mastery experiences. Thus, framing performances in terms of individual student progress might affect students' self-efficacy positively even when the topic is not novel.

In contrast to my significant growth curves (with the exception of the non-significant slope of grade 10 students' self-efficacy for easy tasks), Bernacki et al. (2015) found no significant mean-level changes in students' self-efficacy across four tasks, arguing individual fluctuations (i.e. increases and decreases) masked the variability in their data. I measured self-efficacy across lessons, rather than tasks, and found evidence of both individual variability and mean-level changes (overall increase), for both grade levels. The differences between my study and that of Bernacki et al. (2015) might be related to timing / design. Similar to some of the grade 10 classes in my study, participants in the study by Bernacki et al. worked on equations (2015, p. 105), however there is no information in their paper indicating this topic was new to the students. It is also possible that the differences are related to the larger time lag between measurement occasions in my study, and that the self-efficacy of the participants

in my study grew as an effect of accumulated mastery experiences (while the participants in the study by Bernacki et al. reported their self-efficacy after every single performance experience). Finally, the differences may be related to methodology: Bernacki et al. (2015) used a single item measure of self-efficacy, meaning they were not able to separate measurement error from "real" changes in self-efficacy.

Now moving on to the third key finding: My findings demonstrated there were differences between the growth of grade 6 and grade 10 students' self-efficacy, where younger students' self-efficacy was generally stronger at the beginning of the first lesson, and also changed in a steeper fashion, as compared to those of the grade 10 students. This was true for medium difficulty and hard tasks. Furthermore, slope variance was significant or marginally significant for all task difficulties in the case of the grade 6 students, but not in case of the grade 10 students. Slope variance is indicative of stability of rank-order - if students' self-efficacy grows at significantly different rates, this would presumably cause changes to their rank-order. According to Gottfried et al. (2001), students' rank-order stability might increase as a function of age across motivational domains. My results from research question 3 provide empirical support for age-related differences in stability in the case of mathematics self-efficacy, across much shorter time frames.

The growth curves demonstrated that grade 10 students reached a point of "stability", where additional lesson experiences did little to change their self-efficacy, sooner than the grade 6 students, overall. Inspecting the estimated growth curves in Figure 46 it appears as though grade 10 students' self-efficacy largely stabilised after the first lesson, while grade 6 students' self-efficacy appeared to still be changing after the third lesson. These differences might be due to student age (e.g. cognitive maturation), systematic differences between the lessons of

students in grade 6 and grade 10, or both. Children generally become more adept at assessing their own and other's competence with increasing age (Nicholls, 1978; Stipek, 1981), and as such grade 10 students might have aligned their self-efficacy with their (perceived) performance capability more quickly than grade 6 students (e.g. on the basis of a single lesson experience). As per my discussion in section 4.5.3, previous studies have demonstrated higher stability of competence beliefs with age. Wigfield, Eccles, and Rodriguez (1998) argued students' self-perceptions become "established" during the middle childhood and early adolescent years. The findings by Davis-Kean et al. (2008) indicated a shift in the stability of students' self-concepts after grade 6 (US). My findings indicated changes in the stability of students' self-efficacy still occur between grades 6 and 10, pointing to early intervention (i.e. grade 6, rather than grade 10) in the case of affecting changes to maladaptive self-efficacy.

An alternative explanation for the different shapes of the growth curves is systematic differences between grade 6 and grade 10 students' lessons. It is possible the topics were more novel to the students in grade 6 as compared with grade 10, due to the approach to teaching mathematics in Norway, where later topics build on previous ones ("spiralprinsippet"). While for the phase 2 data collection I requested to be present when the teachers introduced a new topic, I did not include measures of students' perceptions of these topics, and to what degree they were novel to them. Changes in self-efficacy as a function of the degree of novelty of the topic / task will be important to investigate in future studies.

It is also possible the new topics were introduced more quickly in grade 10 than in grade 6, resulting in additional opportunities for novel mastery experiences after the first lesson in grade 6, while there were less such novel experiences available for grade 10 students during lesson two and onwards. Also, social comparisons might be more prevalent in grade 10 versus

grade 6 classrooms, and grade 10 students in Norway have finishing exams at the end of the year, possibly contributing to a focus on normative information. These factors might have lead to different opportunities in regard to both performance experiences and cognitive appraisals of these (i.e. mastery experiences).

Regardless whether differences in self-efficacy growth are a function of age or lesson differences, it appears a focus on normative and / or social standing in the classroom does not support (positive) changes in students' self-efficacy. Alternatively, students' self-efficacy may benefit from a focus on mastery experiences. In contrast to a comparative focus (where only some students can be "better than" the others), individual mastery experiences can serve as a positive source of self-efficacy for all students. I will investigate the relationship between students' self-efficacy and their mastery experiences in study 4 of the thesis.

In terms of key finding four; while my findings regarding the stability of self-efficacy did not indicate large differences between easy, medium difficulty, and hard tasks, the latent growth models indicated marked differences between the mean-levels of these self-efficacy constructs, which are noteworthy even though they were not tested for statistical significance. Unsurprisingly, the intercept of students' self-efficacy was lower when task difficulty was higher, indicating students were more confident about their capabilities on easy, as compared with hard, tasks. Higher task difficulty seemed to also be associated with steeper growth curves, across both groups.

One interpretation is that students' self-efficacy for easy tasks are already high, thus the flat growth might be related to a "ceiling effect", i.e. there is not much room to grow. The descriptives in section 4.3.5 displayed higher values of skew in association with easy items, as

compared with medium difficulty and hard items, in my self-efficacy measure. However, the growth of self-efficacy was steeper for medium difficulty and even more pronounced for hard tasks, which indicates these differences reflect more than a methodological artefact. It is possible that students' self-efficacy is increasingly dependent on personal performance experiences as evidence of capability with increasing task difficulty. In contrast with easy tasks, students are less certain of their capabilities in relation to more difficult, novel, tasks, and thus benefit from task specific feedback in the form of personal performance experiences. This has implications for teachers, and demonstrates the importance of highlighting student progress relative to previous task performance, in order to support students' self-efficacy.

In addition, students might have been particularly uncertain about their ability to succeed on hard tasks prior to the first lesson due to unfamiliarity with task requirements. Students' experiences during the first lesson likely reduced the amount of uncertainty about the requirements of the new mathematical topic, and provided students with mastery experiences of different level tasks. While early mastery experiences seemed particularly important to self-efficacy growth in the case of harder tasks, this added benefit of task experiences to increases in self-efficacy declined across the lessons (as per the flattening growth curves in Figure 46). The low intercept of students' self-efficacy for hard tasks serves to illustrate the kinds of self-doubts students might have toward perceived hard tasks when they lack information from previous experiences, while the relatively steep growth slopes (as compared with medium difficulty and easy tasks) demonstrate that experiences from a single lesson of learning a new topic are able to affect positive changes in these self-doubts, at the group level. Overall, these findings demonstrate that a multidimensional approach to measuring self-efficacy can provide useful substantive insights, and that it is worthwhile to consider task difficulty when investigating students' self-efficacy.

Limitations

There were individual differences in my data, demonstrated through the significant slope variances in relation to the grade 6 students, meaning the growth of students' self-efficacy differed significantly between individuals (thus the grade 6 slope parameters for medium difficulty and hard tasks can be interpreted as the "average" slope, and do not necessarily characterise the growth of all individuals). Furthermore, there were group differences in my data, demonstrated through the ICCs (section 4.4.4.1), which may relate to difference between the classes in terms of parental background, school organisation, or classroom quality. Previous studies (e.g. Hornstra et al., 2013) have reported significant differences in slope between groups of students, considering for instance parental educational level. In the present study I did not have access to individual data such as socioeconomic background, thus I was not able to control for these potential influences. Beyond the issue of generalisability of my group-level findings (to other students and educational contexts), the question of moderators of the process of self-efficacy change is a key issue for future investigations. Another fruitful avenue for further research would be to investigate to what degree individual and group (e.g. classroom) level factors interact with changes in students' self-efficacy, for instance to what degree cognitive appraisal moderates the effect of mastery experiences on changes in students' self-efficacy.

4.5.5 Further implications from study 3

The high rank-order stability found in research question 2 together with the significant mean changes in self-efficacy found in research question 3 demonstrate that while the students' self-efficacy increased (at the group level) across the lessons, the rank-order of students tended to remain stable across these mean changes. Although the differences in stability between grade 6 and 10 were not investigated in research question 2, slope variances from

research question 3 indicated there was more variation in the growth of younger, as compared with older, students' self-efficacy, and more variability in regard to hard tasks, as compared with medium difficulty and easy tasks. Significant variance in growth slope is likely related to more changes of rank-order, i.e. less stability of the self-efficacy construct. Overall, my findings support the theoretical tenets that self-efficacy is a dynamic construct and changes in relation to influential events and shifting circumstances. At the same time, my findings demonstrate the rank-order standing of students' self-efficacy is stable, and most likely particularly so in the case of older students and when students engage with easier tasks. It will be important to formally investigate differences in self-efficacy stability according to student age and levels of difficulty in future studies.

While my findings regarding mean-level changes convey a hopeful message to teachers about the possibility for positive changes across relatively short periods of time, the implications from the high rank-order stability indicate mean-level changes may not be accompanied by changes to the classroom rank-order. It appears as though students who have weaker self-efficacy as compared with their classmates are unlikely to climb on the rank-order, even when their self-efficacy strengthens after engaging with new tasks. While high stability is most likely beneficial for students with strong self-efficacy (given the benefits of optimistic self-efficacy, see section 4.1), it can be problematic for students with low self-efficacy, if they fail to adjust their self-efficacy even at times when this would be reasonable, e.g. after successful performances. As discussed in section 4.5.3, students who have come to see themselves as less capable in mathematics might be prone to look for confirming evidence of this self-perception. For teachers this implies it is important to identify students with weak self-efficacy in mathematics as early as possible, and support positive cognitive appraisals of their performances, as the stability of negative self-perceptions is likely to increase with age.

Students' cognitive appraisals might be related to factors such as classroom climate (e.g. implicit messages about effort and ability), which implies teachers might need to work with a long-term view to support adaptable performance appraisals for all students, particularly those who have a negative view of their own capabilities. Future investigations into the stability of students' cognitive appraisals over time could provide important knowledge for researchers and teachers.

4.5.6 Further limitations of study 3

I discussed some limitations already in relation to each of my findings. I will now raise further limitations not sufficiently captured previously. An important limitation of my study stems from the non-equivalent measurement points of the different classes (see section 4.3.6.2), which violated one of the assumptions of latent growth models. Stemming from this, my findings should not be interpreted in relation to the passing of time (linear), but rather the experience of events (e.g. classroom lessons). In other words, my data is not suited to shed light on the growth of self-efficacy as a function of time passed, but should be interpreted with reference to students' experiences with events (across lessons in mathematics). For instance, it would not make sense to predict on the basis of my findings how much students' self-efficacy changes in the space of any given week, as the time lag varied across my participants, and my study was focused on lessons where a new topic was introduced. However, my findings indicated there is a mean change rate (and a variance around this mean, in the case of grade 6 students) for students' self-efficacy across lessons in mathematics when a new topic was introduced – and that the shape of this slope is largely non-linear. Thus, my findings contribute to our knowledge about how students' self-efficacy changes in relation to experiences of engaging with mathematics, i.e. as a function of perceived experiences/ events, not as a function of time itself.

A further limitation of the present study is that my data does not facilitate investigations into classroom quality or in-depth analyses of the tasks the students engaged with (although an overview of the main task of each lesson was given in Table 28), which will be important directions for future research. Furthermore, grade 10 students in my sample differed to the pool of participants during phase 1, in terms of higher mean scores on national tests (see section 4.3.2.1). It is possible the grade 10 classes that participated during phase 2 differed to other grade 10 students in Norway. Finally, my sample size is modest, my findings are related to Norwegian grade 6 and 10 students, and a Norwegian school context. It will be valuable to investigate the stability and change of students' mathematics self-efficacy with a larger sample, other grade levels and within other socio-cultural contexts.

4.6 Conclusion

I tested a new instrument for measuring students' self-efficacy for learning during lessons in mathematics, developed from my measure of self-efficacy for a national test in mathematics investigated in study 1 of this thesis. I replicated the best-fitting measurement model from study 1, including levels of difficulty as well facets of mathematics. For students in grade 6 and grade 10 alike, this measure was measurement invariant across four lessons in mathematics. Furthermore, I investigated both the rank-order stability as well as mean-level changes of students' self-efficacy across a sequence of lessons in mathematics when students were introduced to a new topic. This is the first study I know of that has investigated self-efficacy change across classroom lessons in mathematics.

My findings demonstrate that the stability of rank-order of students' self-efficacy is highly stable, even as they engage with a new mathematical topic. While significant slope variance in grade 6, and not grade 10, indicated higher stability for older, as compared with younger, students, the proposition that rank-order stability of students' self-efficacy increases with age

should be tested in future longitudinal studies. Furthermore, my findings give empirical evidence of mean-level changes across lessons in mathematics, where the growth of students' self-efficacy was characterised by a sharp initial increase, followed by a flattening or stabilising trend. This was especially typical for self-efficacy for hard tasks, while the growth shape tended to flatten with lower task difficulty. This demonstrates that working on new tasks is particularly salient to students' self-efficacy for hard tasks, and highlights the importance of considering levels of perceived difficulty when investigating students' self-efficacy.

The present study used a micro-analytic approach with a strategic Norwegian sample in order to increase our understanding of changes in self-efficacy across lessons in mathematics when students were introduced to a new topic. Future studies should investigate how Norwegian students' self-efficacy typically develops across classroom learning, in a nationally representative sample. My study is a first step to understanding the processes through which self-efficacy changes. Further micro-analytic classroom studies should include the relationship with mastery experiences over time, in order to increase our understanding of factors related to changes in self-efficacy. Furthermore, covariates that are theoretically assumed to influence the process through which self-efficacy changes should be included.

5 The relationship between self-efficacy and mastery experiences across a sequence of mathematics lessons (study 4)

According to theoretical tenets, mastery experiences are the most important source for students' self-efficacy. Importantly, performance experiences do not influence self-efficacy directly, but are cognitively appraised in regard to, for example, task difficulty and students' ability conceptions. While a handful of longitudinal studies have provided empirical data on the relationship between mathematics self-efficacy and mastery experiences using longitudinal designs (across months or years), there is little empirical evidence for the theoretically proposed process of self-efficacy change at the micro-level (across lessons or learning events). Investigating the reciprocal relationship between mastery experiences and self-efficacy in this time frame is important in order to further our theoretical understanding of the development of mathematics self-efficacy, and to provide practical implications for educators in school.

In the current study, I build on my findings from study 3 by investigating stability and change of students' self-efficacy across a sequence of lessons in mathematics. I also propose a new measure of mastery experiences, developed to match item-by-item, my measure of self-efficacy in study 3. To address the aims of study 4, I used a micro-analytic design with repeated measures of grade 6 (n=81) and grade 10 (n=100) students' self-efficacy and mastery experiences. I also included an adapted measure of students' ability conceptions, as a means to validate my proposed measures of self-efficacy and mastery experiences. My findings provide support for a multidimensional structure of mastery experiences, including three latent levels of difficulty and correlated uniquenesses for facets of mathematics, similar to my previously investigated measure of self-efficacy. Path models with cross-lagged effects illuminated the relationship between self-efficacy and mastery experiences: the dominant direction was from self-efficacy to mastery experiences, but the effects from mastery experiences to gain in self-efficacy were also substantial. Furthermore, construct stability as well as cross-lagged effects seemed to differ according to task difficulty. My proposed measures were validated through significant relationships with an established measure, where students' fixed beliefs about ability were related to their self-efficacy and mastery experiences.

Key implications from my study are that teachers might want to frame task difficulty in terms of individual challenge rather than how difficult tasks are to others, in order to direct students' attention toward the strategies and the effort required to succeed on tasks. Furthermore, in order to affect positive self-efficacy, teachers might be most successful if starting with the least stable construct, students' mastery experiences. If teachers ascertain how students typically appraise their performance experiences, and support adaptive interpretation based on individual accomplishments, this might also strengthen the effect of students' mastery experiences on their self-efficacy. Finally, emphasising effort and learning gains, rather than ability, are key ingredients to build a classroom climate conducive to positive self-efficacy and mastery experiences.

5.1 Introduction

Self-efficacy is important because they are related to a number of positive learning behaviours and outcomes in mathematics, such as learning strategies, perseverance, and choice of career paths (Betz & Hackett, 1983; Zimmerman, 2000). Understanding how mathematics self-efficacy changes and develops during classroom settings is important both to further our understanding of the construct and to inform classroom practice. Mastery experiences are established as the dominant source of self-efficacy through both theory and empirical studies (Byars-Winston et al., 2017; Usher & Pajares, 2008). According to social cognitive theory (Bandura, 1997) students develop their self-efficacy iteratively on the basis of appraisals of their mastery experiences, which highlights the importance of students' mastery experiences during lessons in mathematics.

When students make appraisals of their performance experiences this might include contextual elements, such as the perceived difficulty of a task. While social cognitive theory posits levels of perceived task difficulty might influence students' self-efficacy (Bandura, 1997), investigations on mathematics self-efficacy have rarely focused on this (see study 1). Students' appraisals are likely also influenced by their conceptions of ability. Previous research has provided empirical support for a positive relationship between an incremental view of ability (e.g. ability as malleable) and students' self-efficacy (e.g. Bonne & Johnston, 2016; J. A. Chen & Pajares, 2010), however we know little about the relationship between self-efficacy and ability conceptions for different levels of perceived difficulty.

While some longitudinal studies have provided empirical support for a reciprocal relationship between self-efficacy and mathematics performance (see Talsma, Schüz, Schwarzer, and Norris, 2018; see also study 2, chapter 3, in the present thesis), these studies have generally

included long gaps between measurement occasions, e.g. months or years. While informative in regard to the predictive effects of self-efficacy / performance across long time frames, such longitudinal designs do not shed light on the theoretically proposed process of self-efficacy change (Bandura, 1997), including cycles of mastery experiences and self-efficacy appraisals across hours or days (see e.g. Lindsley et al., 1995). A process approach to studying change can be described as a focus on an "ordered sequence of learning states" (Schmitz, 2006, p. 433), and is important as it gives the possibility to understand how constructs are related over time. Having laid out the importance of investigating the relationship between self-efficacy and mastery experiences in a process perspective and in relation to different levels of task difficulty, in the following section I will present the theoretical background for this study.

5.2 Theoretical background

In the current section I will first briefly present the main theoretical themes of my current study, e.g. the self-efficacy - mastery experiences relationship, task difficulty, and ability conceptions, before I discuss the benefits of a micro-analytic theoretical framework to studying changes in self-efficacy and mastery experiences. In the following sections I will review empirical research regarding the relationship between self-efficacy and mastery experiences over time.

According to the conceptualisation of self-efficacy there are four theoretically proposed *sources* of self-efficacy. The most important source, according to numerous research studies, is mastery experiences (Byars-Winston et al., 2017). Students' previous experiences of success or failure in mathematics form a basis for the formation of subsequent mathematics self-efficacy. Importantly, mastery experiences are students' *interpretations* of experiences rather than their objective performance. As such, performance experiences do not affect individuals' self-efficacy directly, but are interpreted and made sense of by the individual.

This likely has the effect of ensuring relative stability of self-efficacy, where individuals with high self-efficacy are unlikely to lower their beliefs on the basis of a single negative experience, and similarly individuals with low self-efficacy are unlikely to increase these expectations on the basis of a single experience of success (Bandura, 1997). According to the theoretical tenets of self-efficacy (e.g. Bandura, 1997; Klassen, 2004) it is beneficial to be somewhat optimistic – to hold slightly higher self-efficacy than what one actually is capable of. Bandura and Wood (1989) emphasised that optimism is beneficial to performance outcome so long as the quality of performance (e.g. effort and attention) can have a bearing on the outcome, which includes the majority of tasks in school. This might be particularly beneficial to students in Western, as compared to non-Western, cultural groups (Klassen, 2004), which includes the participants in the current study. Optimistic self-efficacy is related to positive learning behaviours such as increased effort and persistence, which are in turn related to positive performance or learning outcomes. This highlights the proposed *reciprocal relationship* between self-efficacy and mastery experiences.

According to Lindsley et al. (1995) the reciprocal relationship between self-efficacy and performance can take the form of "deviation-amplifying" spirals, that are either downward or upward trending. However, overly optimistic self-efficacy is unlikely to lead to continuous increases in performance outcomes. While optimistic self-efficacy can support students to make the most of their learning and performance potentials, students' current knowledge and ability does place limitations on the level of performance they are able to achieve. A central question is the exact meaning of optimism: whether student optimism is related to a belief in their own capability of influencing their learning /performance through the effort they put in (a sense of empowerment or control, which would be related to initiating positive learning behaviours), or if it is related to a positive outcome expectation (a belief in their own fixed

ability, such as 'I am clever', and thus do not need to put much effort in- a belief that would presumably *not* be related to activation of positive learning behaviours). This highlights the importance of how self-efficacy is operationalised, e.g. whether items used for measuring students' self-efficacy are consistent with theoretical tenets.

Over time, it is likely that overly optimistic (or pessimistic) self-efficacy will become more realistic, if individuals receive accurate feedback about their performance (this state is referred to as equilibrium, e.g. Williams & Williams, 2010). Shea & Howell (2000) found that external task feedback was related to "self-corrections" in the self-efficacy / performance relationship, meaning feedback served to interrupt upward or downward trending spirals. In contrast, individuals who did not receive external feedback were less likely to self-correct their self-efficacy, which the authors argued might indicate it was easier to ignore disconfirming evidence if it was self-generated. This illustrates the role of cognitive appraisals, where individuals might sustain overly optimistic or pessimistic self-beliefs in the face of corrective information, such as "objective" task performance.

According to Bandura (1997), students appraise their performance experiences in regard to the perceived difficulty of the task – while optimal challenges are important in order to facilitate experiences of success, succeeding on hard tasks is likely to boost students' self-efficacy to a larger degree than succeeding on easy tasks. Similarly Schweinle and colleagues (Schweinle et al., 2013) argued students' perceptions of task difficulty are important, as persevering and succeeding on challenging tasks is associated with greater learning gains. Previous studies have ascertained that the relationship between task difficulty and competence beliefs is complex, and that individuals rely on a range of sources to assess their performance capability relative to task demands (e.g. Malmberg & Little, 2007; Stajkovic & Luthans,

1998). Furthermore, cognitive maturity (e.g. age) is likely to influence individuals' appraisals of task difficulty as well as their perceptions of competence (Nicholls, 1978; Schweinle, Turner, & Meyer, 2006; Stipek, 1981), where older students are generally more accurate and less optimistic in their self-assessments. Stajkovic and Luthans (1998) included task complexity as a moderator in their meta-analysis of the relationship between self-efficacy and work-related performance (objective measures of task complexity were included in the analysis, which differ from the levels of perceived difficulty as included in my study). Stajkovic and Luthans (1998) found that task complexity significantly moderated the relationship between self-efficacy and performance: the relationship was strongest in relation to the least complex tasks (correlation $G_{(r+)} = .53$), and progressively weakened with increasing task complexity (correlations $G_{(r+)} = .38$ and $G_{(r+)} = .24$ for medium and high complexity, respectively). The authors argued these differences might disappear with repeated performance trials, as individuals became more familiar with the complex tasks. However, they emphasised these relationships were not straightforward, listing ability conceptions and skill acquisition as potential factors contributing to the task complexity/self-efficacy relationship over time.

Ability conceptions are individuals' beliefs about ability as either fixed and more or less inherited (*entity beliefs*), or as changeable and dependent on learning (*incremental beliefs*) (Dweck, 2008; Dweck & Leggett, 1988). Ability conceptions are associated with the types of achievement goals students pursue (demonstrating performance versus learning), and whether students engage in self-handicapped behaviours (Rhodewalt, 1994). Furthermore, Butler (2000) demonstrated that ability conceptions were associated with the sources students attend to during self-appraisal, where incremental beliefs were associated with attending to temporal feedback (e.g. comparisons with own previous performance), while entity beliefs were

associated with normative comparisons (e.g. comparisons with others). Similarly, Bandura (1997) argued students' ability conceptions play a role in how they appraise performance feedback (Bandura, 1997). Students with high entity beliefs are likely to interpret "failures" as evidence of low (fixed) ability, and thus lower their self-efficacy when they encounter problems or make mistakes. In contrast, students who view ability as malleable (high incremental beliefs) are likely to see "failures" as part of the learning process or indicative of effort or task difficulty, and maintain positive self-efficacy despite encountering difficulties. Previous studies have provided empirical support for the relationship between students' ability conceptions and their self-efficacy in mathematics (e.g. Bonne & Johnston, 2016) and science (e.g. J. A. Chen & Pajares, 2010). Bonne and Johnston (2016) found significant associations between students' mathematics self-efficacy and their incremental (r = .25) and their entity (r = -.31) domain-general beliefs. Davis et al. (2011) adapted the original instrument by Dweck (2000) to a mathematics context, and found a significant association between students' mathematics self-efficacy and their mathematics incremental beliefs (r = .26). Younger students are more likely to hold incremental beliefs as compared with older students (Dweck & Leggett, 1988). Furthermore, several researchers have emphasised ability conceptions are not fixed traits and might differ across domains as well as with cognitive maturity (e.g. Bråten & Strømsø, 2005; Buehl & Alexander, 2001; Muis, 2004).

A micro-analytic approach to investigating self-efficacy change

There are aspects of the key theoretical propositions from social cognitive theory that cannot fully be investigated within a cross-sectional or even longitudinal (across long time gaps) framework, but requires a focus on changes across shorter time frames, in other words a micro-analytic approach (see e.g. Bernacki et al., 2015). According to theoretical tenets, the process of self-efficacy change can be seen as a sequence of states, including cycles / spirals

of mastery experiences and self-efficacy appraisals across hours or days, as per Lindsley et al. (1995) and Shea & Howell (2000). This reciprocal relationship is not proposed to be straight forward (e.g. with direct influences) but influenced by for instance students' cognitive appraisal. For instance, performance experiences influence self-efficacy through cognitively appraised mastery experiences, and self-efficacy does not influence performance directly, but through influencing their emotions, thoughts and behaviours (see section 1.3.3). While studies within domains such as management behaviour have investigated upward and downward trending spirals of self-efficacy and performance over a short time frame (e.g. Lindsley et al., 1995; Shea & Howell, 2000), empirical evidence for the proposed reciprocal relationship between the mathematics self-efficacy and performance of school-age children is scarce, across long or short time frames. In the next section I will present the studies I have identified and discuss their contributions to knowledge.

5.2.1 Reciprocal effects between mathematics self-efficacy and mastery experiences

The theoretically proposed *mutual relationship* between mathematics self-efficacy and mastery experiences has been investigated empirically in only few longitudinal studies I am aware of. In my own reciprocal effects study (see study 2, section 3.5.1, in which participants were followed up within a 1-year time-frame) I found positive associations between self-efficacy and national test scores at the first time point as well as one year later (correlations ranging from .15 to .31 in year 2, controlling for the scores in year 1), and a positive effect from national test scores in year 1 on self-efficacy in year 2 (beta ranging from .19 to .28), but not in the other direction (e.g. non-significant effects from self-efficacy in year 1 to national test scores in year 2). Talsma et al. (2018) used the data from 11 studies (six of which considered mathematics) to conduct a meta-analysis on the relationship between self-efficacy and performance, and to investigate in a cross-lagged panel model whether one of the unidirectional effects models or a reciprocal effects model fit better. They included in their

analysis studies investigating academic self-efficacy and performance, with children or adults. Overall, they found support for reciprocal effects, where the effect from performance on self-efficacy was three times as strong ($\beta = .21$) as the effect from self-efficacy on performance ($\beta = .07$), however the effect from self-efficacy on performance was significant only for adults, not children.

The study by Talsma et al. (2018) highlighted some moderating effects relevant to this study. First, larger effects were associated with studies using measurements consistent with theoretical tenets of self-efficacy. Previous reviews (e.g. Klassen & Usher, 2010; Pajares, 1996b) have emphasised the importance of correspondence between the conceptualisation and instrumentalisation of self-efficacy constructs. The number of studies that met the inclusion criteria set by Talsma et al. (2018) (mainly that the study was consistent with recommendations from self-efficacy theory) was very low (only 11 out of 347 articles), highlighting this on-going challenge in self-efficacy research. Talsma et al. (2018) found also that studies with longer time lags between measurement occasions lead to weaker cross-lagged effects, as compared to studies with shorter time lags. Finally, Talsma et al. (2018) found that the effect from self-efficacy on performance was not supported in the case of children. Only two of the 11 studies included by Talsma et al. (2018) focused on school settings, e.g. including children as participants (Bonne & Johnston, 2016; Lewis et al., 2012). Although both these studies included repeated measures of self-efficacy and performance in mathematics (across two and three waves, respectively, over the course of a single year), neither investigated the relationship between self-efficacy and performance specifically in their own analyses. I will now discuss a handful of longitudinal studies I have identified that did investigate school-age children's mathematics self-efficacy and mastery experiences /

performance over time, which for unknown reasons were not included in the Talsma et al. (2018) meta-analysis.

In two studies Hannula and colleagues investigated Finnish school students' mathematics self-beliefs and performance across several years. Hannula, Maijala, & Pehkonen (2004) investigated the dominant direction between Finnish grade 5 (n = 101) and grade 7 (n = 90) students' mathematics *self-confidence* and achievement, as measured one and a half years apart. While their main finding suggested students' self-confidence influenced performance, for older females there was an effect also from performance on self-confidence, supporting a reciprocal relationship. To measure students' self-beliefs the authors used 10 items from the self-confidence subscale of the Fennema-Sherman Mathematics Attitudes scales (Fennema & Sherman, 1976). Some of the items included in the Hannula et al. (2004) study seem congruent with the theoretical tenets of self-efficacy theory as exemplified in Klassen and Usher (2010) (e.g. *I am sure that I can learn mathematics*), while some items appear to differ conceptually (e.g. *Mathematics is my weakest school subject*).

Hannula, Bofah, Tuohilampi, and Metsämuuronen (2014) investigated 3502 students' self-efficacy and achievement at grade 3, 6, and 9, again with an aim to determine the dominant direction of the (reciprocal) effects between the constructs. In contrast to the 2004 study, Hannula et al. (2014) found the effect from performance on self-beliefs ($\beta = .34$) was stronger than the effect in the opposite direction ($\beta = .16$). The difference in magnitude was reduced in the cross-lagged effects from grade 6 to grade 9, where the effect of achievement on self-efficacy ($\beta = .30$) was only slightly higher than the effect from self-efficacy on achievement ($\beta = .26$). Both self-efficacy and achievement were stable across time, achievement more so than self-efficacy. The authors did not report whether the reported differences in magnitude

were tested for statistical significance. To measure self-efficacy Hannula et al. (2014) used a four-item "mathematics self-efficacy scale" (example item *Mathematics is an easy subject*). The example item for self-efficacy differs in face validity to items commonly used to measure the construct, as it is an evaluative statement regarding the difficulty of mathematics as a subject, and does not refer to students' confidence they are able to *do* mathematics. As such, the contribution of both of the above studies by Hannula and colleagues to the knowledge about self-efficacy and achievement over time is uncertain.

In contrast to the long time lags commonly used to investigate self-efficacy over time (see Study 3, section 4.2.1), Bernacki, Nokes-Malach, and Aleven (2015) used a micro-analytic design to investigate how learners' self-efficacy related to their problem-solving performance in algebra. 9th grade students in USA solved problems in an intelligent tutoring system ("Cognitive Tutor Algebra"), where sequential prompts were embedded in the program to elicit students' self-efficacy after receiving (immediate and automated) feedback on their performance during learning. A total of four self-efficacy prompts were given across the learning unit, and the students responded to a self-efficacy prompt after the completion of every fourth algebra problem in the unit, e.g. every 5 minutes on average (according to personal communication I had with the first author). In a path model Bernacki et al. (2015) found that previous performances predicted students' self-efficacy, but not the other way around. However, correlational analyses did indicate relationships between self-efficacy and subsequent performance indicators, where increases in self-efficacy predicted help-seeking behaviour, performance, and learning. Bernacki et al. (2015) used a single item (*How confident are you that you could solve a math question like this one in the future?*) to measure students' self-efficacy. While a single item approach may be valid so long as it is

unambiguous, it is not optimal, because among other things researchers cannot take measurement error into account in the analyses.

While the above studies all included direct measures of performance, rather than students' self-reported mastery experiences, Phan (2012) investigated the relationship between students' mastery experiences and the growth of their self-efficacy over the course of one year. In contrast with theory and empirical studies (e.g. Usher & Pajares, 2008), Phan found a negative relationship between these constructs (β = -.26). In order to explain this unexpected finding, Phan questioned whether grade 3 and 4 students had enough personal learning experiences in mathematics to formulate their self-efficacy. This seems unlikely, as previous studies have established positive relationships between mastery experiences and mathematics self-efficacy for 3rd grade students in France (r=.62; Joët, Usher, & Bressoux, 2011), grade 6-8 students in the US (r=.77; Usher & Pajares, 2009), and between mastery experiences and science self-efficacy for grade 5-8 students in the US (r=.55; Britner & Pajares, 2006). One alternative explanation for the surprising finding by Phan (2012) is a faulty conceptualisation of the mastery experiences measure, which was adapted from a measure that had failed to achieve construct validity in previous studies (see Usher & Pajares, 2009). The example item given (*I always get good grades in mathematics*) is not dissimilar to items commonly used to measure self-concept in mathematics (see Bong & Skaalvik, 2003), and a negative relationship between intercept of self-concept and slope of self-efficacy is not a surprising finding (see for instance study 3, section 4.4.4.1, where I found a significant covariance of -.33 between grade 6 students' self-efficacy intercept and slope for medium difficulty tasks).

The above studies provide mixed evidence for the proposed reciprocal relationship between students' mathematics self-efficacy and mastery experiences. The meta-analysis by Talsma et

al. (2018) found the dominant effect was from performance to self-efficacy, and the effect from self-efficacy to performance was non-significant in the case of children. This is in line with my own reciprocal effects study, where I found significant effects from grade 8 to 9 students' national test scores to self-efficacy, but not the other way around (see study 2, section 3.5.1). Bernacki (2015) similarly found significant effects from grade 9 students' performance to their self-efficacy, while effects from students' self-efficacy to performance were not detected in the path model analysis. The two studies by Hannula and colleagues gave empirical support for a dominant effect from self-efficacy to performance (2004), and from performance to self-efficacy (2014) respectively, however their operationalisation of self-efficacy differed with main theoretical tenets. Similarly, the negative relationship between mastery experiences and self-efficacy reported by Phan (2012) is likely due to the author's conceptualisation of the mastery experiences construct. In light of the sparse evidence concerning students' mathematics self-efficacy and mastery experiences over time, I will briefly comment on some relevant studies that focused on other domains or included older participants.

5.2.2 Reciprocal effects in related domains

Reciprocal relationships across long time-gaps

Investigations into the reciprocal relationship between self-efficacy and performance in domains other than mathematics are also scarce (in comparison to the vast amount of cross-sectional studies), despite the fact that a reciprocal relationship has been supported in some studies such as in Caprara (2011) and Hornstra, van der Veen, Peetsma et al. (2013). Caprara et al. (2011) investigated the relationship between academic self-efficacy and academic performance. They used a staggered, multiple cohort design, including 412 Italian junior high school students as they progressed from age 13 (7^{th} grade) to age 19 (13^{th} grade). Academic self-efficacy was measured at ages 13 and 19. The authors used 15 items to measure two

domains of self-efficacy: to master different subject/curricular areas (example item *How well do you do in mathematics?*), and to self-regulate their own learning (example item *How well can you study when there are other interesting things to do?*). At the end of junior high school (age 14), the authors used a composite academic performance measure including scores for four different school subjects, assigned by the teachers. At the end of senior high school (age 19), the authors used students' final grades, as self-reported by the students. The authors specified autoregressive and cross-lagged reciprocal effects between academic self-efficacy and performance, and found that the stability parameter for self-efficacy (β = .30/.31, males/females respectively), was comparable to the stability of their performance scores (β = .32/.36). Furthermore, the cross-lagged path estimates between self-efficacy and performance scores supported a reciprocal effects model, where the effects from self-efficacy on performance (β = .21/.19 for junior high school grades and β = .20/.22 for senior high school grades) were comparable in magnitude to the effect from junior high school performance on students' self-efficacy in senior high school (β = .22/.22).

Hornstra et al. (2013) investigated the relationship between changes in academic self-efficacy and achievement in reading comprehension from grade 3 to grade 6, among 722 Dutch students. Using latent growth modelling, the authors found a significant relationship between the growth of self-efficacy and the growth of reading achievement (β = .52). Furthermore, higher initial reading achievement predicted slightly steeper self-efficacy growth (β = .01). The authors did not investigate the dominant direction of the relationship between the growth of self-efficacy and mastery experiences (e.g. the growth parameters were correlated, rather than one predicted by the other).

Reciprocal relationships across sequences or cycles

While I have identified only one study that included a micro-analytic approach to investigating the relationship between school-age children's mathematics self-efficacy and performance over time (Bernacki et al., 2015), studies in other domains have focused on the relationship between older individuals' self-efficacy and performance across sequences or cycles. Bandura and Wood (1989) investigated the reciprocal relationship between self-efficacy and performance when college students (N= 60) made decisions concerning managerial design in a simulation study. The authors found that perceptions of controllability of the organisations (easy or hard to control) predicted the college students' self-efficacy, and interacted with the changes in self-efficacy over time, that is, whether participants' self-efficacy increased or decreased across trials (the simulation study included 18 trials in total: self-efficacy was measured after 6, 12, and 18 trials). In a path analysis the authors found support for a reciprocal relationship, where participants' previous performance predicted their self-efficacy (β = .45 and .24 across two blocks of trials), and their self-efficacy also predicted their subsequent performance (β = .29 in the third trial block).

In a more recent study including business students' self-efficacy and task performance on novel manufacturing tasks (N=148), Shea & Howell (2000) found that after an initial decrease in self-efficacy after the first trial, self-efficacy and task performance built on each other in an upward trending spiral (the participants completed four self-efficacy / performance cycles within 2 hours). However, the growth in individuals' performance rates slowed over time, which is unsurprising considering the participants were engaging with new tasks. The stability of self-efficacy between trials remained similar across time (autoregressive paths ranging from β = .63 to β = .71), while individuals' task performance seemed to increase in stability across the trials (autoregressive paths at β = .36, β = .48, and β = .72 between each of the

respective trials). Furthermore, while the effect of performance on self-efficacy was relatively stable across time (β = .22, .19, and .18, respectively), the effect of self-efficacy on performance decreased "dramatically" after the two first trials (β = .30, .30, .11, and .04). This might be a reflection of how knowledge and ability places limits on performance, and that e.g. increased effort (due to optimistic self-efficacy) might have a larger bearing on performance outcomes during initial skill acquisition, than during later stages of learning.

5.2.3 Theoretical model and research questions

Having reviewed empirical evidence regarding the relationship between self-efficacy and mastery experiences in regard to student learning of mathematics as well as other domains, in this section I present the theoretical model and research questions for my present study. On the basis of theoretical tenets and empirical evidence, I propose a theoretical model where students' mathematics self-efficacy is reciprocally related with their mastery experiences, as according to Figure 47. While previous studies have demonstrated significant auto-regressive and reciprocal effects of self-efficacy and mastery experiences across varying time spans, this model highlights the process of self-efficacy change. The process of self-efficacy change includes cycles or sequences of mastery experience events and self-efficacy appraisals, e.g. across hours or days, rather than months or years.

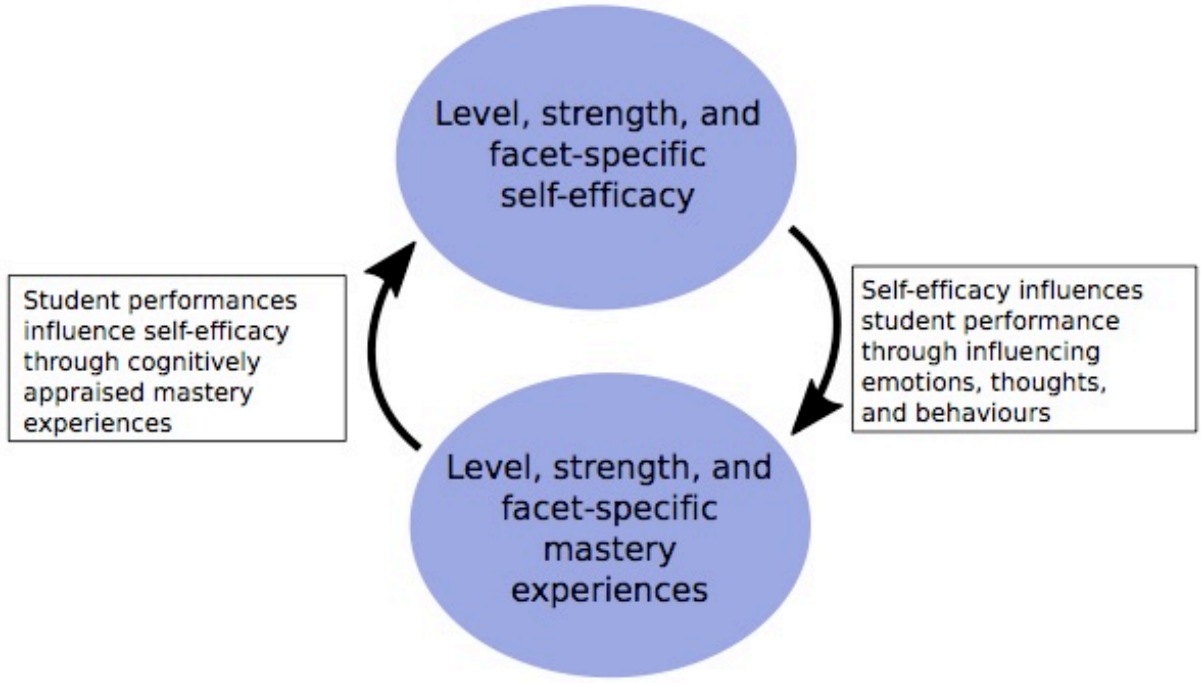


Figure 47: Theoretical model for study 4

I include in this model students' self-efficacy for easy, medium difficulty, and hard tasks, and propose that perceived task difficulty may differentially predict the relationship between self-efficacy and mastery experiences. Furthermore, I propose that students' self-efficacy and mastery experiences are influenced also by their ability conceptions, where higher beliefs in fixed ability predict lower self-efficacy and mastery experiences. Students' grade level predicts their ability conceptions, self-efficacy and mastery experiences due to cognitive maturity and increasing normative comparisons. Accordingly, older students are more likely to hold fixed beliefs of ability as compared with younger students, while younger students are likely to have stronger self-efficacy and mastery experiences as compared with older students.

Research gap

While correlational and longitudinal studies have provided empirical support for a reciprocal relationship between self-efficacy and mastery experiences, there are to the best of my knowledge no studies that investigated the relationship between students' mathematics self-

efficacy and their mastery experiences across a sequence of regular lessons in mathematics. I found one previous study employing a micro-analytic approach to investigating mathematics self-efficacy and performance (Bernacki, Nokes-Malach, & Aleven, 2015), however this study was in the context of an automated algebra tutor rather than a typical classroom setting. Furthermore, I know of no studies that investigated how students' conceptions of ability may relate to their self-efficacy for different levels of task difficulty.

Research questions

In order to address the above gaps in the research I created a new measure of mastery experiences and I employed a micro-analytic design to investigate the following research questions:

(1) What is the structural validity of the proposed mastery experiences constructs, which include facets and levels of mastery experiences?
(2) What is the relationship between self-efficacy and mastery experiences across a sequence of regular lessons in mathematics?
(3) What is the relationship between students' self-efficacy, mastery experiences, and ability conceptions?

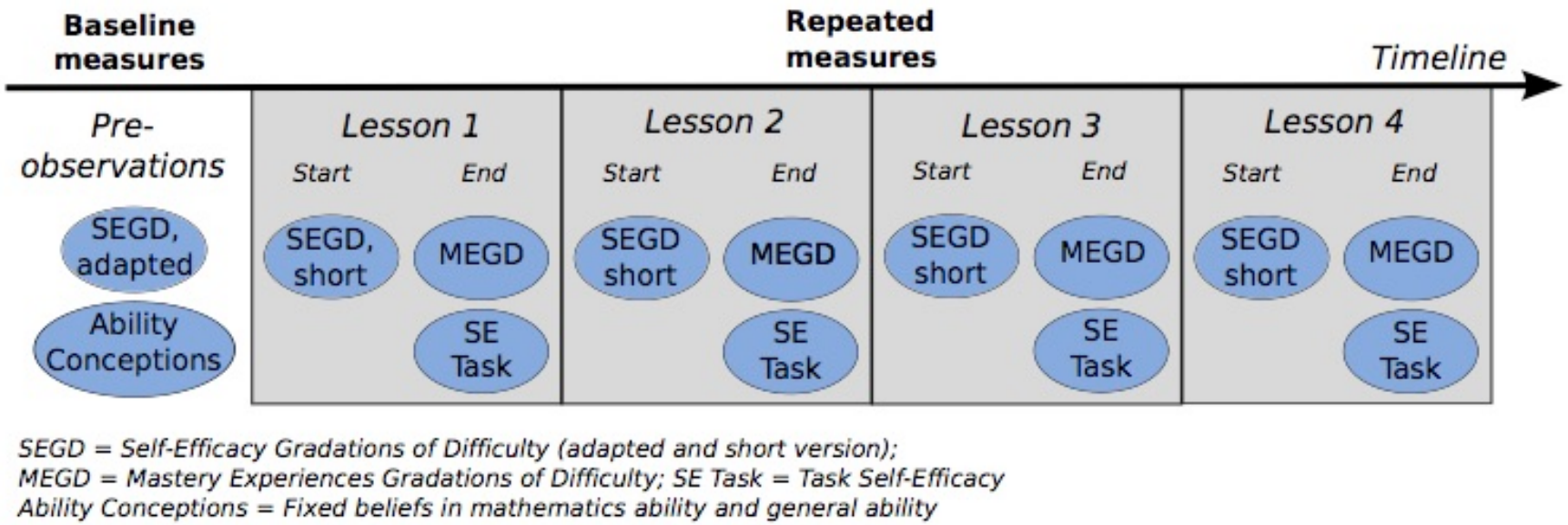
In terms of research question 1 I expect a model including latent levels of difficulty will provide the best fit to the data, similar to the best-fitting model of self-efficacy in study 3. For research question 2 I expect to find a significant reciprocal relationship between students' mastery experiences and their self-efficacy, in line with social cognitive theory (Bandura, 1997). On the basis of previous research, where the dominant direction was from performance to self-efficacy, I expect the dominant direction of the effect to be from students' mastery experiences to their self-efficacy. In terms of research question 3 I expect students' ability conceptions will be significantly related to their self-efficacy and mastery experiences. Previous research has provided empirical evidence for a self-efficacy / ability conceptions

relationship (e.g. Bonne & Johnston, 2016; Davis et al., 2011), thus such a relationship would support the validity of my measure of self-efficacy. While I know of no previous study that investigated the relationship between students' ability conceptions and their mastery experiences, I expect a similar relationship to that of students' self-efficacy, due to the close relationship posited between students' mastery experiences and self-efficacy, as well as previous research indicating that ability conceptions moderate the types of sources students attend to when forming their self-appraisals (Butler, 2000).

5.3 Methodology

5.3.1 Design

In order to address the research questions posed in section 2.3, I designed a short-term longitudinal study. The overall design for this study is the same as for study 3, and is illustrated visually in Figure 48 (see also section 4.3.1). The study targets a relevant time frame for self-efficacy change as well as the processes through which self-efficacy and mastery experiences are related, in accordance to the theoretical propositions from social cognitive theory and empirical findings.



SEGD = Self-Efficacy Gradations of Difficulty (adapted and short version);
MEGD = Mastery Experiences Gradations of Difficulty; SE Task = Task Self-Efficacy
Ability Conceptions = Fixed beliefs in mathematics ability and general ability

Figure 48: Design of study 4

In my micro-analytic design students in four Norwegian grade 6 classes (n=81) and five grade 10 (n=100) classes (mean age 11 and 15, respectively) responded to questionnaires for each of a sequence of 3-4 lessons, when they were introduced to a *new topi*c in mathematics (see study 3, section 4.3.2.1, for details on participant selection, which were identical in studies 3 and 4). The requirement for a new topic was chosen because students' self-efficacy is most likely to change when students engage with new tasks (Usher & Pajares, 2008). As well as students' self-efficacy, measured at the beginning of each lesson (see study 3), in this investigation I drew also on students' mastery experiences, reported at the end of each lesson. A recent meta-analysis by Talsma et al. (2018) reported the effect from self-efficacy on performance was not supported in the case of children. Furthermore, they found that longer time lags were associated with weaker cross-lagged effects between self-efficacy and mastery experiences (Talsma et al., 2018). Including repeated measures of school age students' mastery experiences as well as self-efficacy in my study enabled investigations into this theoretically proposed relationship operationalised at a micro-analytic level.

5.3.2 Measures

In order to investigate the three research questions, I used five measures. Two of these, which were the short and the long versions of a multidimensional measure of students' self-efficacy (the SEGD), were presented in Study 3 (see section 4.3.5). In addition, I created two new measures for the present study: a single item Task self-efficacy measure, and a mastery experiences measure (details of both presented below). Finally, I adapted a measure of ability conceptions from Dweck (2008) for the purposes of the present study (also presented below).

Each of the measures were piloted, and the repeated measures were piloted several times. Similar to the SEGD, short version, the Task self-efficacy measure and the mastery experiences measure (MEGD) were piloted three times with small groups of students in

grades 6 to 8 (see section 4.3.5 for details regarding piloting). The ability conceptions measure was piloted once with a small group of grade 4 students, to ensure the items would make sense to the majority of students in grade 6 and up. During the pilot study the students filled out the questionnaires while I circulated between them, and they were encouraged to ask questions or to comment on items that did not make sense to them. Afterwards, I conducted cognitive interviews with each student, asked them what their interpretation was of each item, and to explain why they chose to respond as they did.

Task self-efficacy measure (Appendix O)

This is a one-item instrument measuring students' self-efficacy to solve mathematical tasks. It refers specifically to the main task of the preceding lesson, thus it was adapted for each class and each lesson. The students responded to the single item task self-efficacy measure at the end of each lesson, indicating their confidence on a 0 to 10 scale. Descriptive scores for the Task self-efficacy measure for each lesson are presented in Table 73. The descriptives indicate that grade 6 students were more confident in regard to performing similar future tasks as compared to the grade 10 students, reflected in their higher mean scores. Both student groups had lower mean self-efficacy scores at the end of the last lesson, as compared to the end of the first lesson. However, for both groups the mean scores were higher at the end of lesson 3 than at the end of lesson 2, thus the change does not appear to be linear.

Table 73: Descriptives for the Task Self-Efficacy measure

	Item	grade 6							grade 10						
		n	missing	M	SD	Range	Skew Z	Kurt Z	n	missing	M	SD	Range	Skew Z	Kurt Z
1	Lesson 1: How certain are you that you can solve similar tasks?	63	18	9.25	1.23	5	-6.64	6.28	85	15	7.94	2.21	10	-4.76	2.59
2	Lesson 2: How certain are you that you can solve similar tasks?	73	8	8.86	1.65	6	-4.67	**0.96**	83	17	7.41	2.72	10	-4.11	**0.54**
3	Lesson 3: How certain are you that you can solve similar tasks?	58	23	9.00	1.70	10	-10.17	22.16	83	17	7.72	2.08	7	-2.19	**-1.56**
4	Lesson 4: How certain are you that you can solve similar tasks?	62	19	8.94	1.45	5	-4.36	**1.54**	69	31	7.04	2.61	10	-2.56	**-0.38**
	Valid N (listwise)	41							51						

Absolute value of skew/ kurt higher than recommended cut-off (Kline, 2011, p. 63)

Mastery Experiences Gradations of Difficulty (MEGD) Questionnaire (Appendix N)

This instrument is a multidimensional measure developed for the present study, to measure students' mastery experiences of learning, problem solving, and self-regulation in mathematics. The MEGD is modelled on the short version of the SEGD, and each of the 9 mastery experiences items are matched with a self-efficacy item, in regard to both facet-specificity and difficulty. Like the SEGD, short version, the MEGD refers specifically to the topic of the previous lesson. This was activated through an introductory statement (*For each of these questions, think about the lesson on <algebra> which has just ended*) presented at the beginning of the questionnaire, as well as questions linked to each of the three facets of mathematics (e.g. *How true is it that you succeeded in learning what the teacher taught you about <algebra> in today's lesson?*), followed by statements of varying gradations of difficulty, to which the students responded (e.g. *In today's lesson I succeeded in learning everything the teacher taught us*). The word in brackets (here: algebra) was changed for each class, to correspond with the topic of their sequence of lessons. Students responded on a scale from 0 to 10, with the anchors 0 = Not at all true, 5 = Moderately true, and 10 = Completely true. Previously, larger effects between self-efficacy and mastery experiences have been found in studies where measurements are aligned with theoretical tenets, e.g. following recommendations by researchers in self-efficacy (Talsma et al., 2018). In my study I have operationalised the measures of self-efficacy and mastery experiences according to suggestions by Bandura (2006) and in accordance with the main conceptualisations of the constructs. Furthermore, I include students' self-reports of mastery experiences rather than objective performance measures which may differ from their perceptions of success, an important distinction from previous studies (see e.g. Talsma et al., 2018).

Descriptives for the MEGD for each lesson are presented in Tables 74 to 77. These demonstrate that students reported higher scores of mastery experiences for easy, followed by medium, and hard, tasks. While students' self-efficacy scores appear comparable across the two samples, there is a tendency that grade 6 students reported higher levels of mastery experiences, than the grade 10 students (the grade 6 students reported higher absolute mean scores for each item, for each of the four lessons). Across both samples, standard deviations are higher in association with the hard tasks, as compared with the easy tasks. Furthermore, estimates of skew and kurtosis seem to be progressively higher in association with easier tasks. More items have severe skew or kurtosis among the grade 6 than the grade 10 sample. I accounted for the non-normally distributed data by using the robust maximum likelihood (MLR) estimator in MPlus (Lei & Wu, 2007). The Cronbach's alpha for the MEGD measure was .93, .96, .96, and .95 for lessons 1-4, respectively. Considering each item score across the sequence of lessons (see Table 78), grade 6 students reported higher mean scores after lesson 4 as compared to lesson 1 on 6 out of 9 items, while grade 10 students reported consistently lower scores for mastery experiences after lesson 4 than after lesson 1.

Table 74: Descriptives Mastery Experiences, lesson 1

	Item	grade 6							grade 10						
		n	missing	M	SD	Range	Skew Z	Kurt Z	n	missing	M	SD	Range	Skew Z	Kurt Z
In today's lesson I succeeded in learning…															
1	at least a little bit of what the teacher taught us	67	14	9.69	1.26	6	-13.15	23.17	87	13	9.10	2.24	10	-10.74	14.18
2	at least half of what the teacher taught us	67	14	9.66	1.12	5	-10.67	14.61	85	15	8.65	2.74	10	-7.92	6.26
3	everything the teacher taught us	67	14	9.10	1.53	7	-7.15	7.16	85	15	7.51	3.38	10	-4.33	**-0.32**
In today's lesson I succeeded in solving…															
4	all the tasks that seemed easy to me	67	14	9.75	0.88	5	-13.41	27.63	85	15	9.18	2.05	9	-9.80	10.81
5	all the tasks that seemed of medium difficulty to me	67	14	8.91	2.31	10	-8.67	10.87	83	17	8.29	2.84	10	-6.54	3.76
6	all the tasks that seemed hard to me	67	14	7.78	2.94	10	-4.80	**1.68**	83	17	6.93	3.44	10	-3.11	**-1.29**
In today's lesson I managed not to give up…															
7	when I was faced with slightly difficult tasks	66	15	9.27	2.26	10	-10.73	16.00	87	13	9.20	2.09	10	-11.52	17.01
8	when I was faced with medium difficulty tasks	66	15	9.12	2.40	10	-9.96	13.17	84	16	8.60	2.71	10	-7.73	6.36
9	when I was faced with very difficult tasks	67	14	8.43	2.81	10	-6.73	4.96	84	16	7.76	3.31	10	-4.69	**0.25**
	Valid N (listwise)	66							82						

Absolute value of skew/ kurt is higher than recommended cut-off (Kline, 2011, p63)

Note: z-score is larger than 1.96 for all items, *except* those in **bold**

Table 75: Descriptives Mastery Experiences, lesson 2

	Item	grade 6							grade 10						
		n	missing	M	SD	Range	Skew Z	Kurt Z	n	missing	M	SD	Range	Skew Z	Kurt Z
In today's lesson I succeeded in learning…															
1	at least a little bit of what the teacher taught us	74	7	9.46	1.45	6	-9.59	10.84	87	13	9.00	2.26	10	-9.39	10.20
2	at least half of what the teacher taught us	74	7	9.20	1.65	7	-7.91	7.30	87	13	8.48	2.91	10	-7.11	3.97
3	everything the teacher taught us	74	7	8.26	2.44	8	-4.77	**0.94**	87	13	7.11	3.59	10	-3.78	**-1.02**
In today's lesson I succeeded in solving…															
4	all the tasks that seemed easy to me	74	7	9.65	1.08	5	-11.81	18.77	87	13	8.60	2.67	10	-7.60	5.50
5	all the tasks that seemed of medium difficulty to me	74	7	9.09	1.62	6	-6.58	4.66	87	13	7.86	3.09	10	-5.29	**1.43**
6	all the tasks that seemed hard to me	74	7	7.69	2.63	10	-3.68	**0.13**	86	14	6.31	3.60	10	-2.30	-2.14
In today's lesson I managed not to give up…															
7	when I was faced with slightly difficult tasks	74	7	9.46	1.63	10	-14.37	32.92	87	13	8.67	2.75	10	-8.56	7.70
8	when I was faced with medium difficulty tasks	73	8	9.15	1.88	10	-9.93	15.33	87	13	8.16	3.11	10	-6.18	2.43
9	when I was faced with very difficult tasks	74	7	8.16	2.80	10	-5.64	2.65	87	13	7.10	3.71	10	-3.59	**-1.30**
	Valid N (listwise)	73							86						

Absolute value of skew/ kurt is higher than recommended cut-off (Kline, 2011, p63)

Note: z-score is larger than 1.96 for all items, *except* those in **bold**

Table 76: Descriptives Mastery Experiences, lesson 3

	Item	grade 6							grade 10						
		n	missing	M	SD	Range	Skew Z	Kurt Z	n	missing	M	SD	Range	Skew Z	Kurt Z
In today's lesson I succeeded in learning…															
1	at least a little bit of what the teacher taught us	61	20	9.75	0.98	5	-13.14	25.45	85	15	8.96	2.19	10	-9.03	9.83
2	at least half of what the teacher taught us	62	19	9.52	1.34	8	-12.56	28.35	85	15	8.31	2.72	10	-5.71	2.25
3	everything the teacher taught us	61	20	8.77	1.97	10	-7.14	9.54	85	15	7.20	3.61	10	-3.83	**-1.01**
In today's lesson I succeeded in solving…															
4	all the tasks that seemed easy to me	61	20	9.74	0.93	5	-12.08	22.67	85	15	8.92	2.16	10	-8.08	7.72
5	all the tasks that seemed of medium difficulty to me	60	21	9.40	1.32	5	-8.50	10.38	85	15	8.04	2.92	10	-5.02	**0.96**
6	all the tasks that seemed hard to me	60	21	8.00	2.56	10	-5.47	4.11	85	15	6.67	3.66	10	-2.86	**-1.85**
In today's lesson I managed not to give up…															
7	when I was faced with slightly difficult tasks	62	19	9.73	1.01	5	-12.32	21.91	85	15	8.75	2.41	10	-7.54	6.28
8	when I was faced with medium difficulty tasks	61	20	9.56	1.29	6	-11.74	21.11	85	15	8.25	2.81	10	-5.54	**1.80**
9	when I was faced with very difficult tasks	61	20	8.92	2.06	10	-8.29	11.22	85	15	7.18	3.36	10	-3.56	**-1.01**
	Valid N (listwise)	60							85						

Absolute value of skew/ kurt is higher than recommended cut-off (Kline, 2011, p63)

Note: z-score is larger than 1.96 for all items, *except* those in **bold**

Table 77: Descriptives Mastery Experiences, lesson 4

	Item	grade 6							grade 10						
		n	missing	M	SD	Range	Skew Z	Kurt Z	n	missing	M	SD	Range	Skew Z	Kurt Z
In today's lesson I succeeded in learning…															
1	at least a little bit of what the teacher taught us	63	18	9.71	1.11	7	-15.29	39.30	70	30	8.61	2.46	9	-6.17	3.71
2	at least half of what the teacher taught us	63	18	9.52	1.20	5	-9.22	11.82	70	30	7.99	2.90	10	-4.66	**1.22**
3	everything the teacher taught us	63	18	8.44	2.18	9	-5.12	3.11	70	30	6.86	3.61	10	-3.17	**-1.06**
In today's lesson I succeeded in solving…															
4	all the tasks that seemed easy to me	64	17	9.84	0.74	4	-16.22	39.05	69	31	8.55	2.44	10	-5.75	3.70
5	all the tasks that seemed of medium difficulty to me	64	17	9.47	1.20	5	-9.48	13.24	69	31	7.74	3.08	10	-3.77	**-0.15**
6	all the tasks that seemed hard to me	64	17	7.89	2.42	9	-3.83	**0.79**	69	31	6.19	3.56	10	**-1.57**	-2.02
In today's lesson I managed not to give up…															
7	when I was faced with slightly difficult tasks	64	17	9.47	1.93	10	-14.06	30.42	70	30	8.56	2.54	8	-5.55	**1.96**
8	when I was faced with medium difficulty tasks	64	17	9.27	1.82	10	-11.14	20.92	70	30	8.04	2.91	9	-4.00	**-0.24**
9	when I was faced with very difficult tasks	64	17	8.33	2.61	10	-5.33	3.08	70	30	7.09	3.48	10	-2.78	**-1.44**
	Valid N (listwise)	63							69						

Absolute value of skew/ kurt is higher than recommended cut-off (Kline, 2011, p63)

Note: z-score is larger than 1.96 for all items, *except* those in **bold**

Table 78: Mastery Experiences mean scores, lesson by lesson

	Item	grade 6				grade 10		
		L1	L2	L3	L4	L1	L2	L3
In today's lesson I succeeded in learning…								
1	at least a little bit of what the teacher taught us	9.69	9.46	9.75	9.71	9.10	9.00	8.96
2	at least half of what the teacher taught us	9.66	9.20	9.52	9.52	8.65	8.48	8.31
3	everything the teacher taught us	9.10	8.26	8.77	8.44	7.51	7.11	7.20
In today's lesson I succeeded in solving…								
4	all the tasks that seemed easy to me	9.75	9.65	9.74	9.84	9.18	8.60	8.92
5	all the tasks that seemed of medium difficulty to me	8.91	9.09	9.40	9.47	8.29	7.86	8.04
6	all the tasks that seemed hard to me	7.78	7.69	8.00	7.89	6.93	6.31	6.67
In today's lesson I managed not to give up…								
7	when I was faced with slightly difficult tasks	9.27	9.46	9.73	9.47	9.20	8.67	8.75
8	when I was faced with medium difficulty tasks	9.12	9.15	9.56	9.27	8.60	8.16	8.25
9	when I was faced with very difficult tasks	8.43	8.16	8.92	8.33	7.76	7.10	7.18
	Valid N (listwise)	66	73	60	63	82	86	85

Ability Conceptions measure, adapted from Dweck (Appendix P)

I adapted the Ability Conceptions questionnaire developed by Dweck (2008) to measure students' beliefs about mathematics ability and general ability. This measure was developed to investigate people's implicit beliefs about intelligence. It includes eight items, where four of the items are posited to measure entity beliefs (e.g. item 8: *People can learn new things, but they CANNOT really change how smart they are*), and four incremental beliefs (e.g. item 5: *No matter who someone is, they can significantly change how smart they are*). The entity items are negatively worded (see items 1, 4, 6, and 8 in Table 79), while the incremental items are positively worded. As such, the entity items might also load on a method factor, accounting for the negatively worded items (Marsh et al., 2010). The anchors on the 1-6 Likert-type scale are formulated thus that higher scores equal higher disagreement with each

statement. I translated the items to the Norwegian myself (see Appendix Q for the Norwegian version), and adapted four of the items (items 1-4 in Table 49: two entity and two incremental beliefs items), by referring to mathematics abilities specifically. A previous study employed a similar approach, adapting six original items by Dweck to a mathematics context, and achieved an alpha of .75 (Davis et al., 2011). In the current study, my measure of ability conceptions (including all eight items) resulted in an alpha of .84, while the alphas for the four mathematics ability items (.78) and the four general ability items (.77) were slightly lower.

I made two adjustments to the original measure based on piloting experiences: I put the negating words in capitals so that the students would not misinterpret the items by mistake, and I changed the somewhat abstract term "intelligence" with "being smart", which was more familiar for my participants. Looking at the mean scores overall (Table 79), there is a tendency that grade 6 students disagree more with entity items/agree more with incremental items as compared with the grade 10 students, indicating the grade 10 students have stronger fixed beliefs in ability on average. Comparing the items adapted to a mathematics context (items 1-4) with the original general ability items (items 5-8) there is a tendency grade 6 and grade 10 students alike have stronger fixed beliefs in association with general ability as compared to their beliefs about ability in mathematics. Finally, there are larger values of skewness and kurtosis associated with the four items about mathematics ability, compared with the original items from Dweck (2008).

Table 79: Descriptives Ability Conceptions

	Item	grade 6							grade 10						
		n	missing	M	SD	Range	Skew Z	Kurt Z	n	missing	M	SD	Range	Skew Z	Kurt Z
1	People CANNOT change whether they are good at mathematics or not	75	6	4.88	1.25	4	-3.91	**0.28**	95	5	4.67	1.28	5	-4.92	2.19
2	Even how good one really is at mathematics one can change a lot	75	6	1.88	0.87	4	3.61	2.30	93	7	2.25	0.95	5	4.42	4.98
3	No matter who someone is, they can always become better at mathematics if they try	73	8	1.44	0.80	4	9.70	16.88	95	5	1.92	1.02	5	5.70	5.50
4	To be honest, people CANNOT change how good they are at mathematics	73	8	5.19	1.11	5	-5.59	4.27	96	4	4.97	1.15	5	-5.83	4.38
5	No matter who someone is, they can significantly change how smart they are	72	9	1.81	0.78	3	**1.91**	**-1.03**	95	5	2.61	1.13	5	2.41	**0.30**
6	How smart someone is, is something they CANNOT change very much	73	8	4.36	1.40	5	-2.60	**-0.85**	96	4	4.27	1.17	5	-2.06	**-0.75**
7	People can always change how smart they are a lot	74	7	2.12	1.09	4	3.62	**1.32**	94	6	2.73	1.10	4	**1.81**	**-0.41**
8	People can learn new things, but they CANNOT really change how smart they are	74	7	4.59	1.28	5	-3.29	**0.74**	96	4	3.86	1.41	5	**-1.71**	**-1.61**
	Valid N (listwise)	71							89						

Note: z-score is larger than 1.96 for all items, *except* those in **bold**

5.3.3 Analyses

As in Study 3, I used SPSS (version 24) for data management and descriptive analyses, while I used Mplus (version 1.31) for structural equation modelling. In order to account for non-normally distributed data (e.g. estimates of skewness and kurtosis for my mastery experiences measure), I used the robust maximum likelihood (MLR) estimator in Mplus (Lei & Wu, 2007). The MLR estimator treats missing data according to the full-information maximum likelihood approach (Graham & Coffman, 2012), using all available data. I used pairwise deletion in SPSS to produce correlation estimates. Consistent with the previous studies in this thesis, I used the following parameter estimates and cut-offs to determine acceptable model fit: The chi square (χ^2/df <3 acceptable), the Root Mean Square Error of Approximation (RMSEA <.08 acceptable), the Comparative Fit Index (CFI) and the Tucker-Lewis index (TLI) (CFI/TLI > .90 acceptable), and the Standardized Root Mean Square Residual (SRMR<.10 acceptable) (Chiorri, 2014; Moring, Marsh, Nagengast, and Scalas, 2014; Schermelleh-Engel, Moosbrugger, and Müller, 2003). If a minimum of three parameters were "acceptable" I included the model for substantive considerations, inspecting whether parameter estimates were meaningful and within bounds, the size of standard errors, and residual variances. If a model resulted in negative residual variance on a single item, I proceeded to run a series of tests to rule out likely reasons for this, as per the recommendations by Chen et al. (2001). As I discussed in study 1, section 2.3.5, residual variance on more than one item or on latent factors is considered unusual, and I have encountered no recommendations in this regard. Accordingly, I dealt with such results through model re-specification (e.g. rejection). For model comparisons I applied ΔRMSEA (.015), ΔCFI (-.010), and ΔSRMR (.030) (Chen, 2007) as cut-offs. For tests of invariance, I considered ΔCFI (-.010) in addition to the $\Delta\chi^2$.

As for some of my research questions in study 3 I combined students from grade 6 and grade 10 in my analyses. While my reason for including different grade levels in the study was guided by prior research indicating differences in students' self-efficacy according to age (see section 1.3.3), several of the models did not converge when specified with a smaller sample. I chose a design in which following my participants over time was central, rather than focusing on a larger sample. Based on my findings in Study 1, I expect that combining the grade levels did not influence the findings from research question 1 in this study (the structural validity of the MEGD), however it is possible grade level has a bearing on the substantive findings in research questions 2 and 3. I include grade level as a covariate to alleviate for this, however caution must still be taken in the interpretations of my findings, as the group-level findings are likely to "hide" individual variation. Below I describe my analytic strategy to investigate each of my research questions in succession.

5.3.3.1 Analytic strategy research question 1

In order to investigate the structural validity of my proposed mastery experiences measure, I replicated the process applied in research question 1, study 3 (see section 4.3.6.2), where I investigated the structural validity of my measure of self-efficacy (the SEGD, short version). Accordingly, I first inspected the item-level correlation matrix, at each of the four time points. I then proceeded to specify a series of plausible empirical models. The empirical models for my mastery experiences measure are presented in Figures 49 to 54. These models mirror those investigated for my self-efficacy measure (see section 4.3.6.2), as the mastery experiences measure was created to include similar multidimensionality (level, strength, and facet-specific mastery experiences) to my measure of self-efficacy. In the following figures ovals (in blue) represent latent factors, while the rectangles (1-9) represent manifest items. Single-headed arrows represent factor loadings. Double-headed arrows between factors represent factor correlations, while double-headed arrows between items (1-9) represent

correlated uniquenesses (CUs) for corresponding facets of mathematics or levels of perceived difficulty.

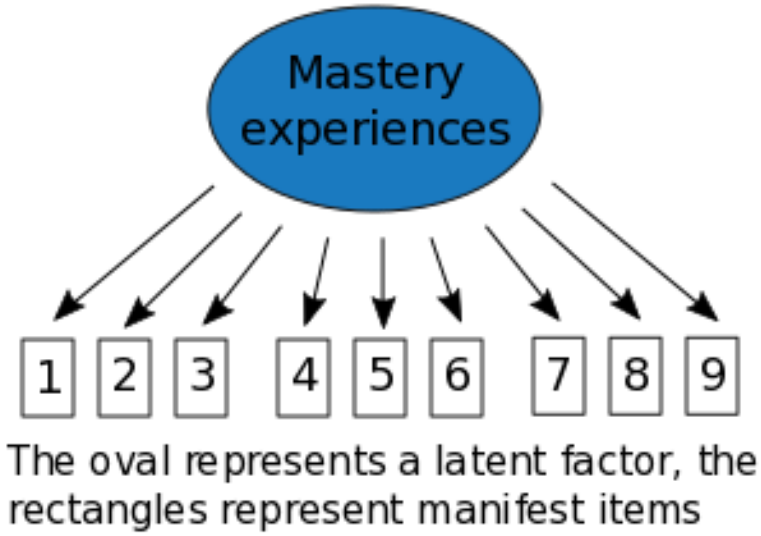


Figure 49: Unidimensional Model (Model 0)

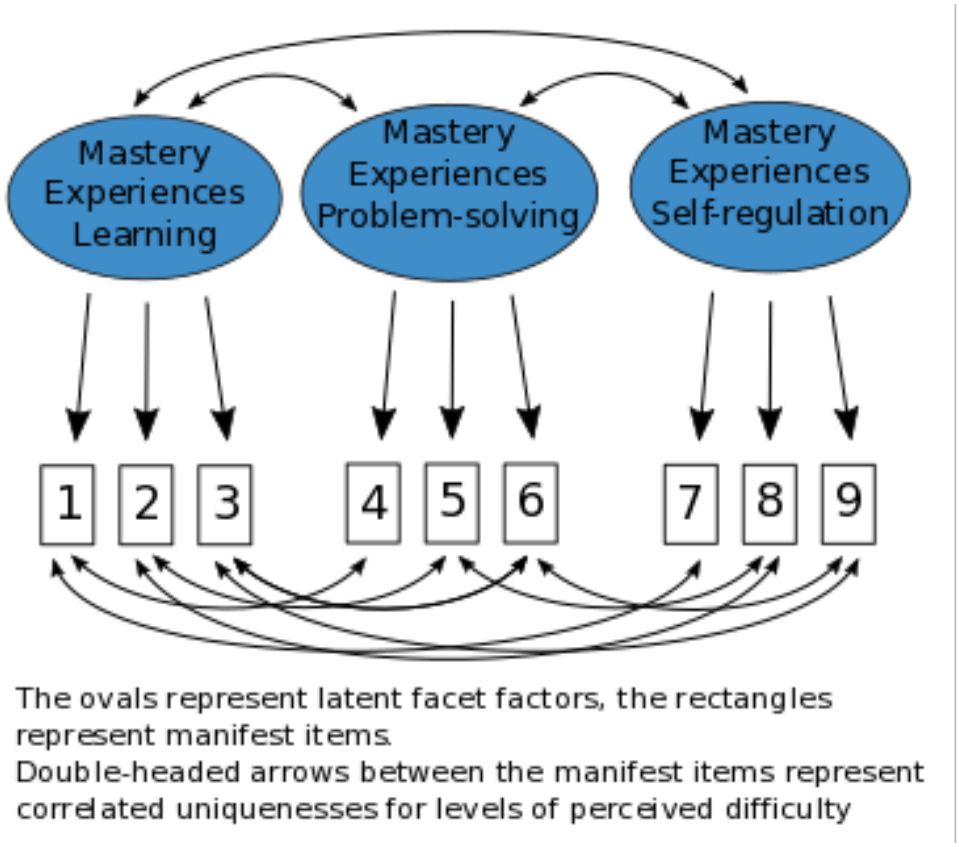
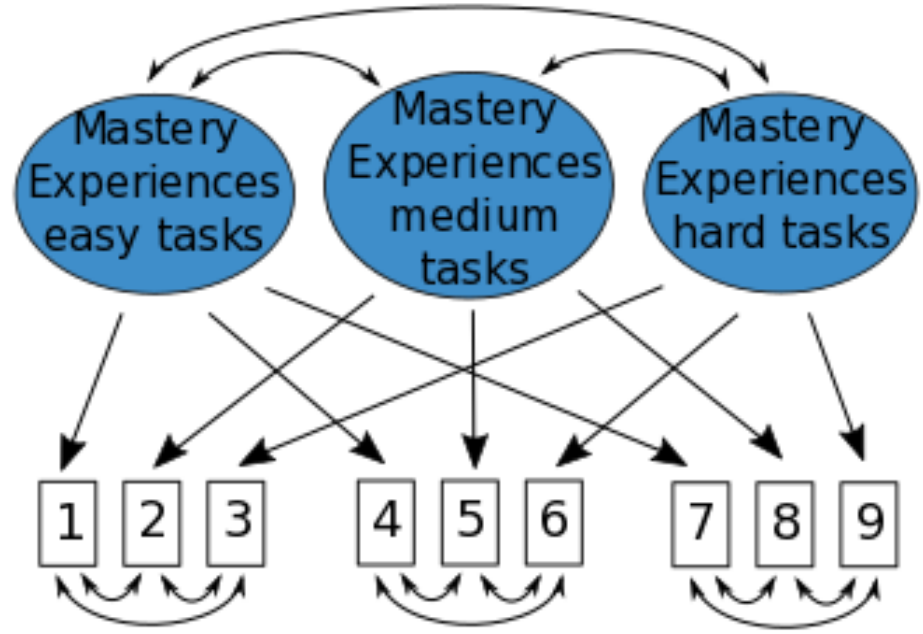
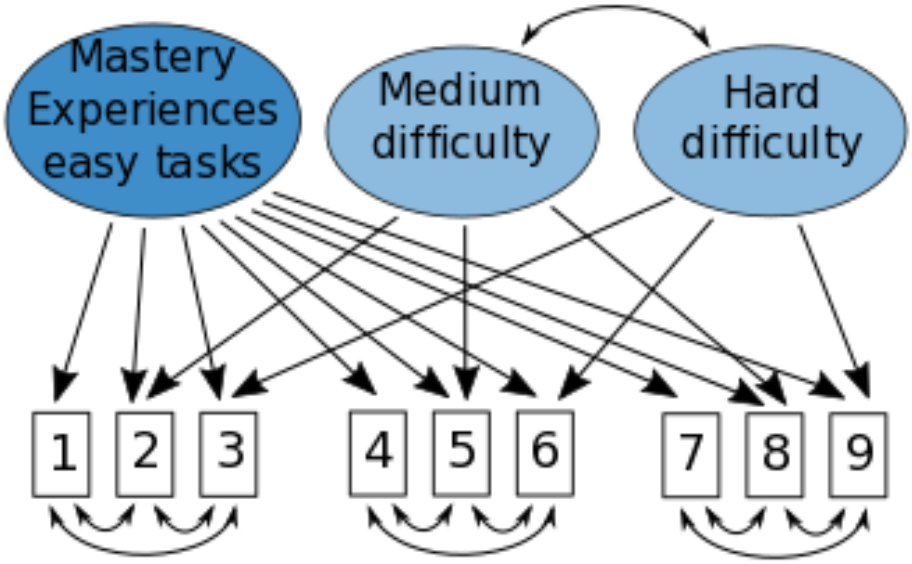


Figure 50: Latent Facets Model (Model 1)



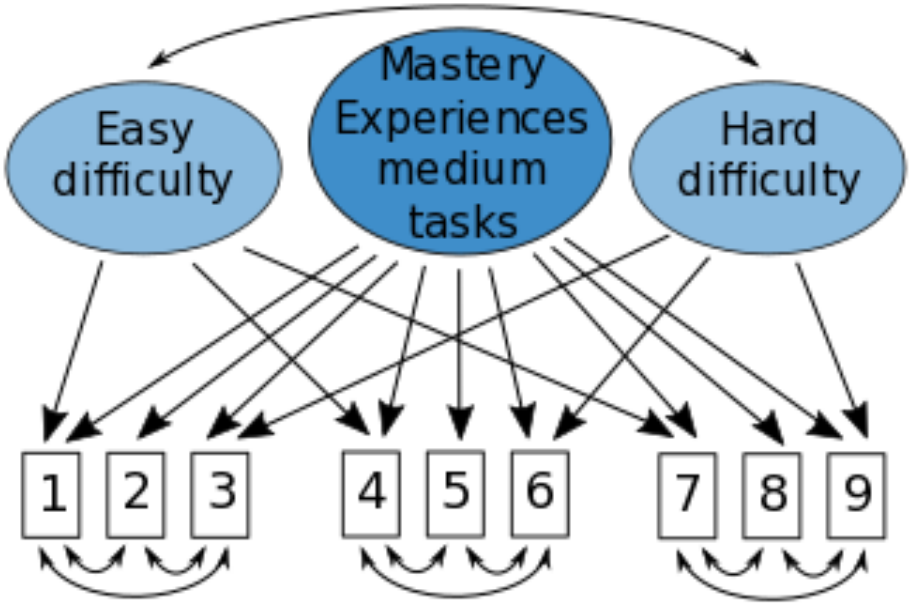
The ovals represent latent level factors, the rectangles represent manifest items
Double-headed arrows between manifest items represent correlated uniquenesses for facets of mathematics

Figure 51: Latent Levels Model (Model 2)



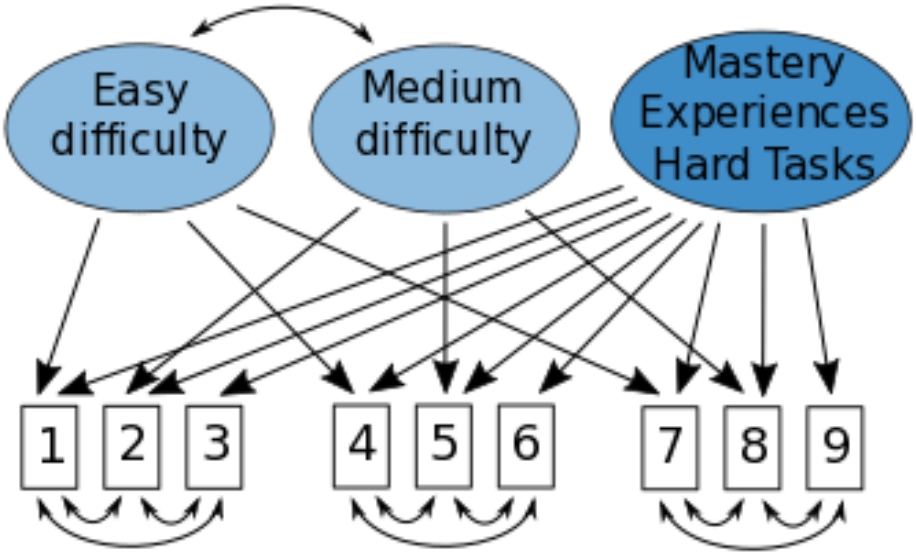
The ovals represent latent factors, the rectangles represent manifest items
Double-headed arrows between manifest items represent correlated uniquenesses for facets of mathematics

Figure 52: Bi-factor Model, easy tasks (Model 3)



The ovals represent latent factors, the rectangles represent manifest items
Double-headed arrows between manifest items represent correlated uniquenesses for facets of mathematics

Figure 53: Bi-factor Model, medium difficult tasks (Model 4)



The ovals represent latent factors, the rectangles represent manifest items
Double-headed arrows between manifest items represent correlated uniquenesses for facets of mathematics

Figure 54: Bi-factor Model, hard tasks (Model 5)

When I had established the best fitting model for Mastery Experiences across the four time points (configural invariance), I proceeded to investigate weak and strong measurement invariance through applying equality constraints. Given the complexity of a multidimensional measurement model, within a repeated measures design, I investigated measurement invariance for each of my proposed level constructs (easy, medium difficulty, and hard tasks) separately.

5.3.3.2 Analytic strategy research question 2

To investigate the relationship between students' self-efficacy and mastery experiences, I first inspected the correlation matrices for my new repeated measures (the MEGD and the Task self-efficacy measure), to see whether a simplex model of change might fit for each of them. The simplex change structure describes a process where the influence of a construct at a single time point *t*, to any proceeding time points, $t + n$, are fully mediated through the adjacent time point, $t + 1$ (Little, 2013, p. 184) (see also section 4.3.6.2). For the multidimensional mastery experiences I included three latent level constructs in the correlation matrix, based on the measurement model from research question 1, and mirroring also the latent level constructs from the SEGD, short version (see study 3, section 4.4.2). I then investigated the stability of students' mastery experiences through formulating three autoregressive simplex models, for easy, medium difficulty, and hard tasks. Similarly, I formulated an autoregressive simplex model to investigate the stability of students' Task self-efficacy. The stability of students' self-efficacy was investigated in study 3, where a simplex change process fit the data well. I proceeded to specify autoregressive models including cross-lagged paths of students lesson self-efficacy and mastery experiences, as well as their Task self-efficacy. Then, I applied equality constraints on the stability and cross-lagged paths, to investigate differences across the lessons and also enable comparison of construct stability and cross-lagged effects.

5.3.3.3 Analytic strategy research question 3

Finally I included students' ability conceptions as a covariate in the models from research question 2, in order to investigate the relationship between students' self-efficacy, mastery experiences, and ability conceptions. While I investigated the factor structure, invariance, and stability over time for each of my measures using item-level analyses, I used parcelling technique in the path models with cross-lagged paths in research questions 2 and 3, due to the complexity of these models. While parcelling is some times debated as researchers may run the risk of misinterpreting the "true" factor structure of a measure (Bandalos, 2002), it can also be useful when the aim is not to test measurement models, but to investigate substantive relationships between latent constructs (Little et al., 2002; see also section 4.3.6.2). When I specified my models at the item level, some of the parameter estimates were not meaningful (e.g. constructs established as highly stable appeared not to be stable at all). I decided parcelling was a reasonable approach in order to investigate the substantive relationships in research questions two and three of the present study, where the parcels were created on the basis of the previously established factor structures of my measures (forming latent level constructs for self-efficacy as well as mastery experiences).

5.4 Results

In the following sections I present the findings for each of the research questions posted. I first present the results pertaining to the factor structure of the new mastery experiences measure (RQ1), in order to answer the substantive research question building on the measures of self-efficacy (the factor structure was investigated in study 3) and mastery experiences (RQ2). In order to answer the last research question, I first present results pertaining to the factor structure of my adapted measure of students' ability conceptions, before presenting my findings regarding the relationship between students' conceptions of ability, self-efficacy, and mastery experiences (RQ3).

5.4.1 RQ1: The structural validity of my mastery experiences measure

5.4.1.1 Correlation matrices

As a first step to investigating the factor structure of the new measure of mastery experiences, I inspected the correlation matrix for each time point (see Tables 80 to 83). Overall, the correlation matrices indicate students' mastery experiences are influenced by both facet-specificity (as indicated by cross-correlation within each facet), and levels of difficulty (as indicated by higher correlations between adjacent levels of difficulty). This supports a multidimensional structure of this measure, as found for the SEGD in Studies 1 and 3 (see sections 2.5.1 and 4.5.2). All non-significant cross-correlations are in association with items of different facets, and different levels of difficulty (with 3 exceptions in the case of item 7, and two exceptions in the case of item 3, across all four lessons). The non-significant correlation estimates are all in association with the grade 6 sample. Among these item 7 (*I managed not to give up when I was faced with slightly difficult tasks*) has the highest number of non-significant estimates (15 of 32 possible), followed by item 3 (*I succeeded in learning everything the teacher taught us;* 9 of 32 possible). Overall, the grade 10 correlations were of a higher magnitude as compared with the grade 6 correlation estimates. All grade 10 correlations were significant and stronger than .50, while in the case of the grade 6 estimates there were several non-significant intercorrelations at each time point, and the lowest magnitude was at .08.

Table 80: Correlation matrix Mastery Experiences, lesson 1

	Item	1	2	3	4	5	6	7	8	9
1	I succeeded in learning at least a little bit of what the teacher taught us	1	.93**	.77**	.84**	.75**	.61**	.60**	.60**	.52**
2	I succeeded in learning at least half of what the teacher taught us	.84**	1	.88**	.85**	.79**	.69**	.59**	.65**	.61**
3	I succeeded in learning everything the teacher taught us	.40**	.68**	1	.75**	.77**	.75**	.49**	.61**	.66**
4	I succeeded in solving all the tasks that seemed easy to me	.41**	.47**	.31**	1	.82**	.65**	.69**	.67**	.61**
5	I succeeded in solving all the tasks that seemed of medium difficulty to me	.31*	.42**	.50**	.46**	1	.85**	.57**	.61**	.61**
6	I succeeded in solving all the tasks that seemed hard to me	.30*	.39**	.52**	.41**	.88**	1	.44**	.56**	.67**
7	I managed not to give up when I was faced with slightly difficult tasks	.48**	.41**	.12	.63**	.23	.20	1	.92**	.72**
8	I managed not to give up when I was faced with medium difficulty tasks	.52**	.47**	.18	.61**	.27*	.25*	.97**	1	.89**
9	I managed not to give up when I was faced with very difficult tasks	.43**	.43**	.23	.50**	.33**	.41**	.80**	.86**	1

** Correlation is significant at the 0.01 level (2-tailed); * correlation is significant at the 0.05 level (2-tailed).
Grade 6 estimates are displayed below the diagonal, grade 10 above

Table 81: Correlation matrix Mastery Experiences, lesson 2

	Item	1	2	3	4	5	6	7	8	9
1	I succeeded in learning at least a little bit of what the teacher taught us	1	.92**	.73**	.91**	.85**	.68**	.88**	.84**	.68**
2	I succeeded in learning at least half of what the teacher taught us	.87**	1	.84**	.87**	.88**	.76**	.83**	.85**	.73**
3	I succeeded in learning everything the teacher taught us	.61**	.78**	1	.77**	.85**	.89**	.74**	.81**	.81**
4	I succeeded in solving all the tasks that seemed easy to me	.42**	.53**	.47**	1	.92**	.73**	.91**	.90**	.75**
5	I succeeded in solving all the tasks that seemed of medium difficulty to me	.53**	.75**	.78**	.72**	1	.85**	.84**	.89**	.79**
6	I succeeded in solving all the tasks that seemed hard to me	.40**	.57**	.74**	.42**	.78**	1	.67**	.76**	.78**
7	I managed not to give up when I was faced with slightly difficult tasks	.20	.23	.13	.12	.17	.08	1	.95**	.79**
8	I managed not to give up when I was faced with medium difficulty tasks	.37**	.40**	.31**	.24*	.43**	.32**	.84**	1	.91**
9	I managed not to give up when I was faced with very difficult tasks	.39**	.47**	.43**	.15	.49**	.59**	.55**	.74**	1

** Correlation is significant at the 0.01 level (2-tailed); * correlation is significant at the 0.05 level (2-tailed).
Grade 6 estimates are displayed below the diagonal, grade 10 above

Table 82: Correlation matrix Mastery Experiences, lesson 3

	Item	1	2	3	4	5	6	7	8	9
1	I succeeded in learning at least a little bit of what the teacher taught us	1	.88**	.70**	.84**	.74**	.59**	.78**	.75**	.64**
2	I succeeded in learning at least half of what the teacher taught us	.56**	1	.92**	.83**	.85**	.77**	.76**	.80**	.73**
3	I succeeded in learning everything the teacher taught us	.53**	.44**	1	.71**	.82**	.82**	.64**	.74**	.73**
4	I succeeded in solving all the tasks that seemed easy to me	.48**	.60**	.19	1	.85**	.68**	.83**	.75**	.69**
5	I succeeded in solving all the tasks that seemed of medium difficulty to me	.77**	.59**	.48**	.54**	1	.90**	.79**	.84**	.75**
6	I succeeded in solving all the tasks that seemed hard to me	.61**	.35**	.51**	.24	.77**	1	.65**	.74**	.78**
7	I managed not to give up when I was faced with slightly difficult tasks	.70**	.81**	.29	.65**	.61**	.47**	1	.92**	.77**
8	I managed not to give up when I was faced with medium difficulty tasks	.67**	.78**	.15	.53**	.59**	.46**	.93**	1	.91**
9	I managed not to give up when I was faced with very difficult tasks	.38**	.53**	.22	.30*	.67**	.67**	.61**	.68**	1

** Correlation is significant at the 0.01 level (2-tailed); * correlation is significant at the 0.05 level (2-tailed).
Grade 6 estimates are displayed below the diagonal, grade 10 above

Table 83: Correlation matrix Mastery Experiences, lesson 4

	Item	1	2	3	4	5	6	7	8	9
1	I succeeded in learning at least a little bit of what the teacher taught us	1	.90**	.71**	.89**	.83**	.63**	.89**	.84**	.68**
2	I succeeded in learning at least half of what the teacher taught us	.81**	1	.91**	.83**	.87**	.70**	.87**	.90**	.75**
3	I succeeded in learning everything the teacher taught us	.38**	.70**	1	.66**	.76**	.69**	.75**	.82**	.72**
4	I succeeded in solving all the tasks that seemed easy to me	.57**	.67**	.44**	1	.85**	.68**	.82**	.82**	.68**
5	I succeeded in solving all the tasks that seemed of medium difficulty to me	.58**	.87**	.73**	.73**	1	.83**	.83**	.86**	.82**
6	I succeeded in solving all the tasks that seemed hard to me	.56**	.63**	.69**	.35**	.68**	1	.65**	.72**	.78**
7	I managed not to give up when I was faced with slightly difficult tasks	.35**	.23	.24	.19	.12	.15	1	.94**	.79**
8	I managed not to give up when I was faced with medium difficulty tasks	.58**	.55**	.26*	.27*	.44**	.39**	.68**	1	.87**
9	I managed not to give up when I was faced with very difficult tasks	.48**	.48**	.37**	.14	.45**	.58**	.45**	.83**	1

** Correlation is significant at the 0.01 level (2-tailed); * correlation is significant at the 0.05 level (2-tailed).
Grade 6 estimates are displayed below the diagonal, grade 10 above

5.4.1.2 Factor structure

For each time point I tested the same six empirical models as for the measure of self-efficacy (see study 3, section 4.3.6.2, and Figures 3 to 8 in section 5.3.3.1). Across all time points (see Tables 84 to 87), the Unidimensional Model (Model 0) resulted in inacceptable fit indices, considering the a priori determined cut-offs (see section 5.3.3). Model 1, the Latent Facets Model, resulted in acceptable fit indices for lessons 2 – 4, however this model also resulted in negative residual variances on several items, at each time point. As I discussed in section 5.3.3 the occurrence of more than one "Heywood case" is rare, with little existing guidance on what to do. In line with study 1, section 2.3.5, I decided not to accept models with more than one negative residual. Models 2-5, including the Latent Levels Model and the three Bi-factor Models, all resulted in acceptable fit indices and meaningful parameter estimates, with acceptable standard errors and no out-of-bounds estimates (e.g. negative residuals). When considering $\Delta\chi^2$, the more parsimonious Latent Levels Model (Model 2) fit best for lessons 1, 3, and 4, while the Bi-factor Easy Model (Model 3) was best fitting for lesson 2. When considering ΔSRMR the Latent Levels Model was equivalent or better fitting for all lessons, save lesson 1 (where the Bi-factor Medium and Hard Models were better fitting). Finally, the Latent Levels Model was equivalent or better fitting across all lessons when considering ΔCFI. Overall, the Latent Levels Model (Model 2) was clearly the best-fitting measurement model for the MEGD, consistent with the measurement model for the SEGD (see section 4.4.2 in study 3).

Table 84: Results confirmatory factor analyses, Mastery Experiences lesson 1

		Fit indices						*Model comparison**			
	Model	χ^2	*p*	RMSEA	CFI	TLI	SRMR	ΔCFI	ΔSRMR	$\Delta\chi^2_{(\Delta df)}$	*p*
0	Unidimensional model	$\chi^2_{(27)}$= 503.34 χ^2/df= 18.64	*p*<.001	0.34	0.32	0.10	0.13				
1	Latent Facets model, correlated uniquenesses for levels**	$\chi^2_{(15)}$= 63.21 χ^2/df= 4.21	*p*<.001	0.14	0.93	0.84	0.08				
2	Latent Levels model, correlated uniquenesses for facets	$\chi^2_{(15)}$= 21.02 χ^2/df= 1.40	*p*=.136	0.05	0.99	0.98	0.05				
3	Bi-factor Easy model, correlated uniquenesses for facets	$\chi^2_{(11)}$= 17.96 χ^2/df= 1.63	*p*=.083	0.06	0.99	0.97	0.03	-0.001	-0.020	$\Delta\chi^2_{(4)}$= -3.96	*p>.05*
4	Bi-factor Medium model, correlated uniquenesses for facets	$\chi^2_{(11)}$= 12.90 χ^2/df= 1.17	*p*=.300	0.03	1.00	0.99	0.02	0.006	-0.033	$\Delta\chi^2_{(4)}$= -8.12	*p>.05*
5	Bi-factor Hard model, correlated uniquenesses for facets	$\chi^2_{(11)}$= 15.12 χ^2/df= 1.37	*p*=.178	0.05	0.99	0.98	0.02	0.003	-0.033	$\Delta\chi^2_{(4)}$= -5.90	*p>.05*

**Negative residual variance on several items, *Models 3-5 were compared against Model 2

Table 85: Results confirmatory factor analyses, Mastery Experiences lesson 2

		Fit indices						*Model comparison**			
	Model	χ^2	*p*	RMSEA	CFI	TLI	SRMR	ΔCFI	ΔSRMR	$\Delta\chi^2_{(\Delta df)}$	*p*
0	Unidimensional model	$\chi^2_{(27)}$= 342.47 χ^2/df= 12.68	*p*<.001	0.27	0.65	0.53	0.07				
1	Latent Facets model, correlated uniquenesses for levels**	$\chi^2_{(15)}$= 27.08 χ^2/df= 1.81	*p*=.028	0.07	0.99	0.97	0.05				
2	Latent Levels model, correlated uniquenesses for facets	$\chi^2_{(15)}$= 16.64 χ^2/df= 1.11	*p*=.341	0.03	1.00	1.00	0.03				
3	Bi-factor Easy model, correlated uniquenesses for facets	$\chi^2_{(11)}$= 5.06 χ^2/df= 0.46	*p*=.928	0.00	1.00	1.02	0.01	0.002	-0.014	$\Delta\chi^2_{(4)}$= -11.58	*p<.05*
4	Bi-factor Medium model, correlated uniquenesses for facets	$\chi^2_{(11)}$= 9.53 χ^2/df= 0.87	*p*=.573	0.03	1.00	1.01	0.02	-0.001	-0.008	$\Delta\chi^2_{(4)}$= -7.11	*p>.05*
5	Bi-factor Hard model, correlated uniquenesses for facets	$\chi^2_{(11)}$= 13.28 χ^2/df= 1.21	*p*=.275	0.04	1.00	0.99	0.02	-0.001	-0.003	$\Delta\chi^2_{(4)}$= -3.36	*p>.05*

**Negative residual variance on several items, *Models 3-5 were compared against Model 2

Table 86: Results confirmatory factor analyses, Mastery Experiences lesson 3

		Fit indices						*Model comparison**			
	Model	χ^2	*p*	RMSEA	CFI	TLI	SRMR	ΔCFI	ΔSRMR	$\Delta\chi^2_{(\Delta df)}$	*p*
0	Unidimensional model	$\chi^2_{(27)}$= 145.07 χ^2/df= 5.32	*p*<.001	0.17	0.78	0.71	0.06				
1	Latent Facets model, correlated uniquenesses for levels**	$\chi^2_{(15)}$= 17.24 χ^2/df= 1.15	*p*=.305	0.03	1.00	0.99	0.07				
2	Latent Levels model, correlated uniquenesses for facets	$\chi^2_{(15)}$= 19.85 χ^2/df= 1.32	*p*=.178	0.05	0.99	0.98	0.02				
3	Bi-factor Easy model, correlated uniquenesses for facets	$\chi^2_{(11)}$= 22.01 χ^2/df= 2.00	*p*=.024	0.08	0.98	0.93	0.02	-0.011	-0.009	$\Delta\chi^2_{(4)}$= 2.16	*p>.05*
4	Bi-factor Medium model, correlated uniquenesses for facets	$\chi^2_{(11)}$= 14.93 χ^2/df= 1.36	*p*=.186	0.05	0.99	0.98	0.02	0.002	-0.008	$\Delta\chi^2_{(4)}$= -4.92	*p>.05*
5	Bi-factor Hard model, correlated uniquenesses for facets	$\chi^2_{(11)}$= 15.36 χ^2/df= 1.40	*p*=.167	0.05	0.99	0.97	0.02	0.001	-0.002	$\Delta\chi^2_{(4)}$= -4.49	*p>.05*

**Negative residual variance on several items, *Models 3-5 were compared against Model 2

Table 87: Results confirmatory factor analyses, Mastery Experiences lesson 4

		Fit indices						*Model comparison**			
	Model	χ^2	*p*	RMSEA	CFI	TLI	SRMR	ΔCFI	ΔSRMR	$\Delta\chi^2_{(\Delta df)}$	*p*
0	Unidimensional model	$\chi^2_{(27)}$= 151.75 χ^2/df= 5.62	*p*<.001	0.19	0.74	0.65	0.06				
1	Latent Facets model, correlated uniquenesses for levels**	$\chi^2_{(15)}$= 35.84 χ^2/df= 2.39	*p*=.002	0.10	0.96	0.90	0.08				
2	Latent Levels model, correlated uniquenesses for facets	$\chi^2_{(15)}$= 21.86 χ^2/df= 1.46	*p*=.111	0.06	0.99	0.97	0.03				
3	Bi-factor Easy model, correlated uniquenesses for facets	$\chi^2_{(11)}$= 23.96 χ^2/df= 2.18	*p*=.013	0.09	0.97	0.91	0.02	-0.013	-0.012	$\Delta\chi^2_{(4)}$= 2.10	*p>.05*
4	Bi-factor Medium model, correlated uniquenesses for facets	$\chi^2_{(11)}$= 18.05 χ^2/df= 1.64	*p*=.080	0.07	0.99	0.95	0.02	-0.001	-0.005	$\Delta\chi^2_{(4)}$= -3.81	*p>.05*
5	Bi-factor Hard model, correlated uniquenesses for facets	$\chi^2_{(11)}$= 16.38 χ^2/df= 1.49	*p*=.128	0.06	0.99	0.96	0.03	0.003	-0.001	$\Delta\chi^2_{(4)}$= -5.48	*p>.05*

**Negative residual variance on several items, *Models 3-5 were compared against Model 2

Figure 55 displays the Latent Levels Model at lesson 1, for purposes of illustration. The factor loadings and the correlations between the latent level factors for this model differ somewhat, but are comparable across the four lessons.

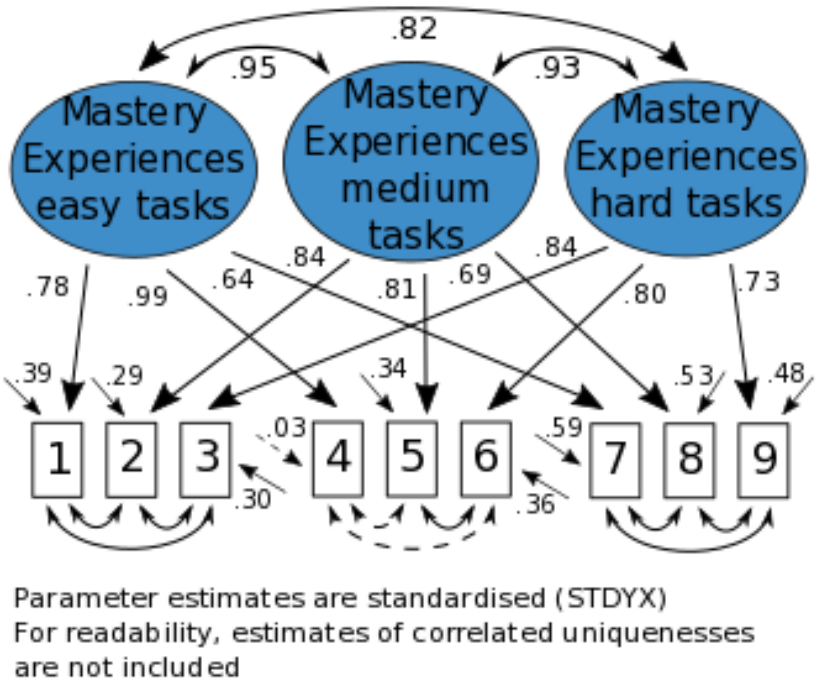
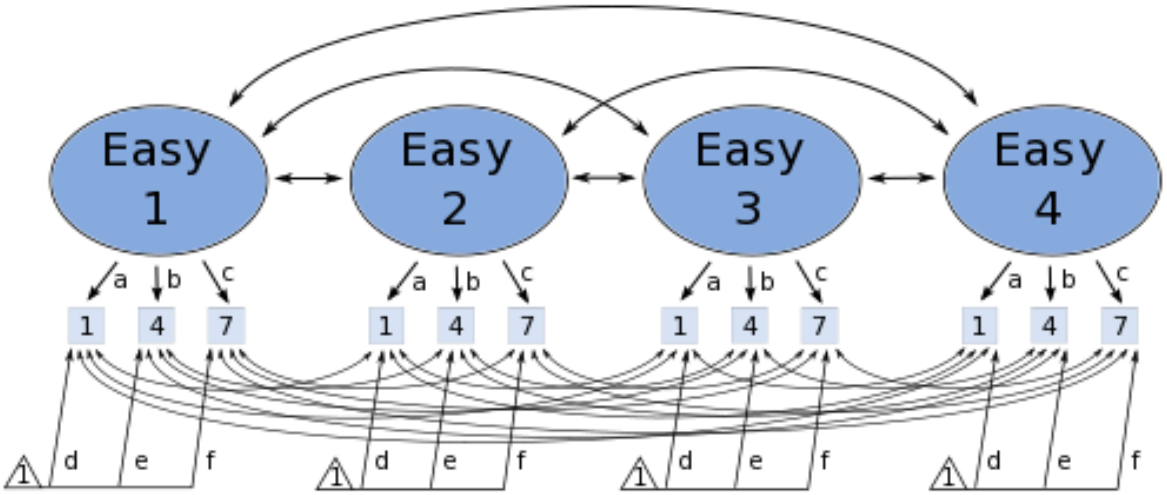


Figure 55: Latent Levels model, lesson 1

5.4.1.3 Measurement invariance of my mastery experiences measure

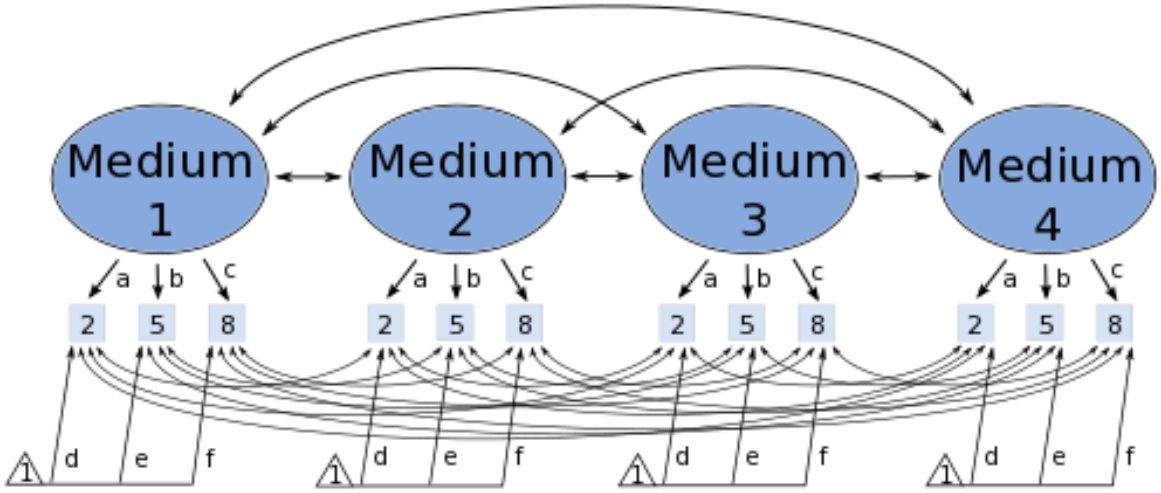
In order to test whether my Mastery Experiences constructs were invariant over time (across lessons 1 to 4), I specified a sequence of models in which increasingly strong invariance assumptions were tested, e.g. equivalent factor loadings (weak invariance) versus equivalent factor structure (configural invariance), and equivalent item intercepts (strong invariance) versus equivalent factor loadings (weak invariance). I first specified the preferred measurement model from section 5.4.1.2 (the Latent Levels Model, Model 2), however when four time points were included this model resulted in several error messages, similar to the Latent Levels Self-Efficacy Model in study 3 (see section 4.4.2.3). The correlation matrix including latent level constructs (see Table 89, section 5.4.2.1) indicates the three latent level

constructs for Mastery Experiences are strongly correlated at each time point, indicating there may be an issue with co-linearity. Accordingly, I proceeded to test for measurement invariance for each of the latent level constructs (Mastery Experiences for easy, medium difficulty, and hard tasks) separately (see Figures 56 to 58).



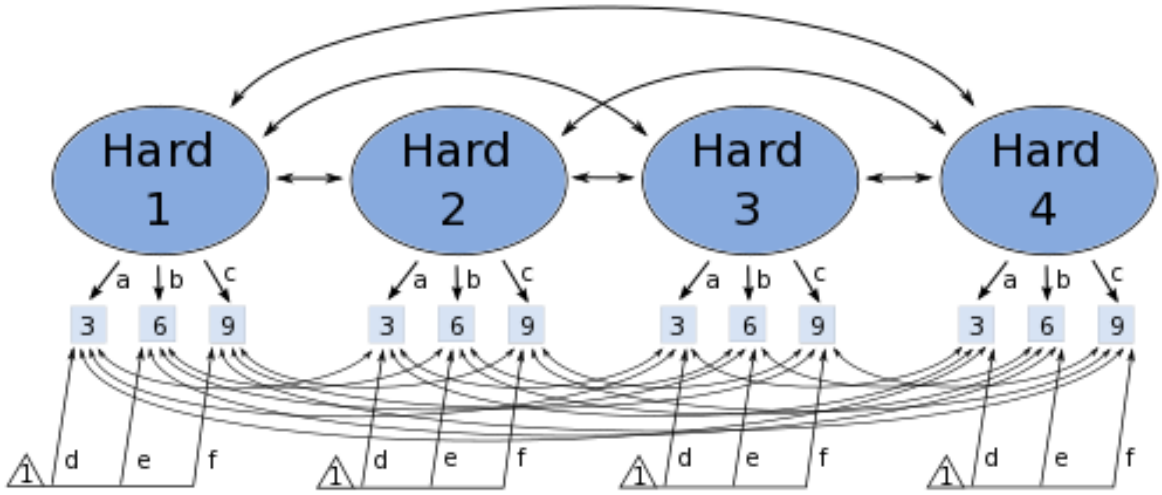
Easy 1, 2, 3, and 4 = Students' mastery experiences on easy tasks
a, b, and c are factor loadings; d, e, and f are manifest (item) means
Weak invariance: parameters a, b, and c are constrained as equal
Strong invariance: parameters d, e, and f are also constrained as equal

Figure 56: Invariance model mastery experiences, easy construct



Medium 1, 2, 3, and 4 = Students' mastery experiences on medium difficulty tasks
a, b, and c are factor loadings; d, e, and f are manifest (item) means
Weak invariance: parameters a, b, and c are constrained as equal
Strong invariance: parameters d, e, and f are also constrained as equal

Figure 57: Invariance model mastery experiences, medium difficulty construct



Hard 1, 2, 3, and 4 = Students' mastery experiences on hard tasks
a, b, and c are factor loadings; d, e, and f are manifest (item) means
Weak invariance: parameters a, b, and c are constrained as equal
Strong invariance: parameters d, e, and f are also constrained as equal

Figure 58: Invariance model mastery experiences, hard construct

The results from the invariance tests are presented in Table 88. In regard to the easy level mastery experiences construct, both factor loadings (weak invariance) and item intercepts (strong invariance) could be constrained as equal across time without significant detriment in

fit. In regard to the medium difficulty and hard level constructs, weak invariance could be applied without detriment to fit, while constraining all item intercepts across the four lessons (strong invariance) resulted in decreased fit considering differences in CFA as well as $\chi 2$. For the medium difficulty model, modification indices indicated problems with item 5 (*In today's lesson I succeeded in solving all the tasks that seemed of medium difficulty to me*), while in the case of the hard level model, the modification indices indicated problems with item 6 (*In today's lesson I succeeded in solving all the tasks that seemed hard to me*), both at lesson 1. When the equality constraints on these items were lifted (parameter *e* in Figures 57 and 58) at lesson 1, both models resulted in fit indices that were not significantly different to the weak invariance model, e.g. the medium difficulty and the hard level constructs were *partial* strongly invariant across the four lessons. As per study 3, section 4.4.2.3, Little (2013, p. 159) argues partial invariance is not problematic as long as it is due to only one of several (a minimum of three) factor items changing meaning over time. I next repeated the steps for invariance testing for the grade 6 and grade 10 students separately, in order to see whether the problem might be assigned to a part of my sample only. In the case of both the medium difficulty and hard level constructs, the models led to error messages in the case of the grade 6 students, possibly due to the small sample size. However, in the case of the grade 10 students, both weak and strong invariance held. This indicates the problems of invariance were related to the grade 6 sample only, and that the two items in question might have changed meaning for some of the younger students between the end of the first and the end of the second lesson.

Table 88: Fit indices Mastery Experiences invariance tests

		Fit indices						*Model comparison****		
	Model	χ^2	*p*	RMSEA	CFI	TLI	SRMR	ΔCFI	$\Delta\chi^2_{(\Delta df)}$	*p*	*pass?*
0	Latent levels model, CU's for facets, no constraints**	$\chi^2_{(438)}$= 858.55	*p*<.001	0.07	0.92	0.88	0.05				
1	Latent easy level model, no constraints	$\chi^2_{(30)}$= 38.01	p=.150	0.04	0.98	0.97	0.04				
1b	Latent easy level model, weak invariance	$\chi^2_{(36)}$= 41.54	p=.242	0.03	0.99	0.98	0.07	0.005	$\Delta\chi^2_{(6)}$= 3.53	*p>.05*	*yes*
1c	Latent easy level model, strong invariance	$\chi^2_{(45)}$= 42.75	p=.568	0.00	1.00	1.01	0.08	0.011	$\Delta\chi^2_{(9)}$= 1.21	*p>.05*	*yes*
2	Latent medium level model, no constraints	$\chi^2_{(30)}$= 49.45	p=.014	0.06	0.97	0.94	0.04				
2b	Latent medium level model, weak invariance	$\chi^2_{(36)}$= 50.92	p=.051	0.05	0.98	0.96	0.05	0.006	$\Delta\chi^2_{(6)}$= 1.47	*p>.05*	*yes*
2c	Latent medium level model, strong invariance	$\chi^2_{(45)}$= 89.56	p<.001	0.08	0.94	0.91	0.08	-0.039	$\Delta\chi^2_{(9)}$= 38.64	*p<.001*	*no*
2d	Latent medium level model, *partial* strong invariance (*item 5, lesson 1)	$\chi^2_{(44)}$= 63.74	p=.027	0.05	0.97	0.96	0.07	-0.006	$\Delta\chi^2_{(8)}$= 12.82	*p>.05*	*yes*
2e	Latent medium level model, weak invariance (grade 10 only)	$\chi^2_{(36)}$= 47.03	p=.103	0.06	0.98	0.96	0.06				
2f	Latent medium level model, strong invariance (grade 10 only)	$\chi^2_{(45)}$= 63.41	p=.036	0.07	0.97	0.95	0.06	-0.013	$\Delta\chi^2_{(9)}$= 16.38	*p>.05*	*yes*
2	Latent hard level model, no constraints	$\chi^2_{(30)}$= 39.52	p=.115	0.04	0.99	0.97	0.05				
2b	Latent hard level model, weak invariance	$\chi^2_{(36)}$= 42.00	p=.227	0.03	0.99	0.99	0.05	0.005	$\Delta\chi^2_{(6)}$= 2.48	*p>.05*	*yes*
2c	Latent hard level model, strong invariance	$\chi^2_{(45)}$= 62.87	p=.040	0.05	0.98	0.97	0.06	-0.015	$\Delta\chi^2_{(9)}$= 20.87	*p<.05*	*no*
2d	Latent hard level model, *partial* strong invariance (*item 6, lesson 1)	$\chi^2_{(44)}$= 55.21	p=.120	0.04	0.99	0.98	0.06	-0.007	$\Delta\chi^2_{(8)}$= 13.21	*p>.05*	*yes*
2e	Latent medium level model, weak invariance (grade 10 only)	$\chi^2_{(36)}$= 46.05	p=.122	0.05	0.98	0.97	0.05				
2f	Latent medium level model, strong invariance (grade 10 only)	$\chi^2_{(45)}$= 59.45	p=.073	0.06	0.98	0.97	0.06	-0.008	$\Delta\chi^2_{(9)}$= 13.40	*p>.05*	*yes*

* *Item not constrained as equal*

***The Latent Levels Model resulted in an error message, thus each of the latent level structures were investigated separately*

****Each model was compared with the equivalent model with one less constraint*

5.4.1.4 Summary of results for RQ1

I tested a series of plausible empirical models of mastery experiences, on the basis of theory as well as previous findings regarding students' lesson self-efficacy (see study 3, section 4.4.2). Similar to my measure of self-efficacy in study 3, a Latent Levels Model, including mastery experiences for easy, medium difficulty, and hard tasks, was retained as the best-fitting meaningful measurement model. This model included correlated uniquenesses for facets of mathematics. For each of the three levels, I tested measurement invariance over time. Students' mastery experiences for easy level tasks stayed strongly invariant across the four lessons, while in the case of medium difficulty and hard level tasks, students' mastery experiences were partially strong factorially invariant. This was in relation to a single item in the case of both medium difficulty and hard level tasks, which could not be constrained equal at lesson 1 for the grade 6 students. This indicates that the meaning of two items (*In today's lesson I succeeded in solving all the tasks that seemed of medium difficulty to me,* and *In today's lesson I succeeded in solving all the tasks that seemed hard to me*), related to the same facet of mathematics, may have changed meaning for the grade 6 students after the first lesson. It is possible that grade 6 students' experiences during the second lesson changed their perceptions of "succeeding", or of what constituted a "medium difficulty" or "hard" task for them. Overall, tests of invariance indicated my measures of mastery experiences for easy, medium difficulty, and hard tasks stayed invariant across the four lessons, and that it is reasonable to apply these measures to further investigate stability and change over time.

5.4.2 RQ2: The relationship between students' self-efficacy and mastery experiences over time

In order to investigate the reciprocal relationship between students' self-efficacy and their mastery experiences I used four measures: the short and long versions of the

multidimensional measure of self-efficacy (the SEGD), the multidimensional measure of mastery experiences (the MEGD), and the single item Task self-efficacy measure. The factor structure of the two multidimensional measures have been previously investigated (see section 4.4.2 in study 3). In addition, the simplex model of change has been investigated in relation to the multidimensional self-efficacy measure (see section 4.4.3.1, study 3). Presently, I will present autocorrelations and investigate simplex models in relation to the mastery experiences and Task self-efficacy measures, before I include all measures in autoregressive models with cross-lagged paths to test for reciprocal relationships.

5.4.2.1 Stability of the mastery experiences measure

Autocorrelations for the mastery experiences measure

I first inspected the correlation matrix including autocorrelations for each of the latent Mastery Experiences level constructs across the four time points (see Table 89). Consistently within each lesson, for both grade levels, the easy level constructs were most closely associated with the medium level constructs, and least strongly correlated with the hard level constructs. Similarly, the hard level constructs were most closely associated with the medium level constructs, and least closely associated with the easy level constructs. These trends are consistent with the factor structure of the mastery experiences measure, including the three level constructs, established in section 5.4.1.2.

I proceeded to inspect the correlation estimates to consider whether a simplex change structure was likely to fit the data well. According to a simplex model of change I would expect the correlation estimates to be consistently stronger for constructs that are measured closer in time. The correlation estimates in Table 89 are inconclusive – while some autocorrelations indicate a consistently weaker relationship over time (as expected in a simplex process), there are also some instances where lesson 2 and lesson 4 autocorrelations,

for instance, are more strongly correlated than corresponding estimates between lessons 3 and 4, or lessons 2 and 3. Five estimates are non-significant. These are all in relation to grade 6, and three of them include the lesson 4 hard level construct with the lesson 1 easy, medium difficulty, and hard level constructs. This means grade 6 students' mastery experiences with hard tasks in lesson 4 were not associated with their mastery experiences with any level of difficulty task in lesson 1.

Table 89: Correlation matrix Mastery Experiences level constructs, lesson 1-4

	Item	1	2	3	4	5	6	7	8	9	10	11	12
1	Lesson 1: Mastery Experiences easy tasks	1	.93**	.77**	.70**	.71**	.60**	.70**	.70**	.63**	.74**	.68**	.58**
2	Lesson 1: Mastery Experiences medium difficulty tasks	.84**	1	.91**	.68**	.76**	.71**	.65**	.68**	.65**	.73**	.74**	.70**
3	Lesson 1: Mastery Experiences hard tasks	.61**	.86**	1	.57**	.68**	.68**	.53**	.57**	.58**	.65**	.68**	.66**
4	Lesson 2: Mastery Experiences easy tasks	.53**	.46**	.40**	1	.95**	.80**	.81**	.70**	.55**	.79**	.80**	.70**
5	Lesson 2: Mastery Experiences medium difficulty tasks	.48**	.43**	.39**	.85**	1	.90**	.76**	.74**	.66**	.81**	.87**	.79**
6	Lesson 2: Mastery Experiences hard tasks	.31*	.41**	.47**	.59**	.81**	1	.62**	.70**	.74**	.69**	.81**	.80**
7	Lesson 3: Mastery Experiences easy tasks	.44**	.44**	.43**	.66**	.67**	.52**	1	.92**	.77**	.76**	.69**	.64**
8	Lesson 3: Mastery Experiences medium difficulty tasks	.50**	.45**	.37**	.73**	.82**	.67**	.87**	1	.93**	.71**	.69**	.67**
9	Lesson 3: Mastery Experiences hard tasks	.28*	.27*	.36**	.62**	.69**	.75**	.55**	.73**	1	.66**	.67**	.65**
10	Lesson 4: Mastery Experiences easy tasks	.31*	.34*	.32*	.65**	.50**	.36**	.48**	.48**	.32*	1	.94**	.80**
11	Lesson 4: Mastery Experiences medium difficulty tasks	.34*	.29*	.21	.73**	.66**	.49**	.50**	.65**	.52**	.74**	1	.92**
12	Lesson 4: Mastery Experiences hard tasks	.18	.14	.21	.46**	.55**	.56**	.24	.45**	.50**	.54**	.79**	1

**Correlation is significant at the 0.01 level (2-tailed), *correlation is significant at the 0.05 level (2-tailed).
Grade 6 estimates are displayed below the diagonal, grade 10 above

Simplex models of mastery experiences

In order to test the simplex models for mastery experiences, I specified three models on the basis of the strongly /partially strongly invariant model specifications (see section 5.4.1.3). Each of these models include a single autoregressive path between each adjacent measurement point, meaning any relationships between non-adjacent lessons, e.g. lesson 1 and 3, are mediated through the lessons in between, e.g. lesson 2. Table 90 displays the fit indices for the three simplex model specifications (Models 1, 2, and 3), each of which were acceptable or very good, based on my chosen fit indices.

I included equality constraints on the autoregressive paths to investigate whether the stability of the rank-order of the constructs changed significantly over time (Models 1b, 2b, and 3b in Table 90). The results from these models indicated that in the case of the easy and the hard level constructs, the stability paths between different time points were not significantly different (e.g. they could be constrained equal without a significant detriment in fit). In the case of the medium level constructs, the model including equality constraints passes when considering CFA, while when considering $\chi 2$ it fails marginally. However, in the interest of parsimony, and considering also that the $\chi 2$ is often deemed "too sensitive" as a sole indicator (see section 4.3.6), I choose the models with equal stability path for all level constructs (Models 1b, 2b, and 3b), including the medium level difficulty construct.

I proceeded to regress each of the simplex models (with equality constraints for stability) on student grade (Models 1c, 2c, and 3c in Table 18), in order to account for differences between the grade levels. The parameter estimates from these models are displayed in Figures 59 to 61. I chose to display the standardised path estimates here, while in section 5.4.3.6 the unstandardised autoregressive paths for mastery experiences will be included in a cross-

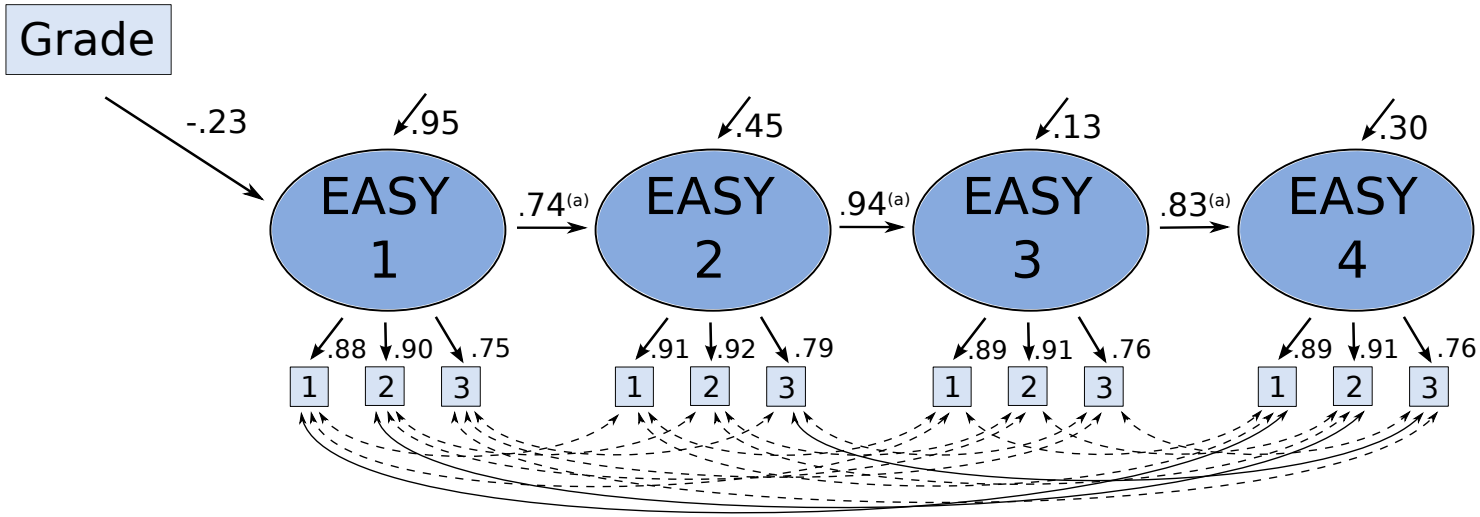
lagged path model with the two self-efficacy constructs. Figures 59 to 61 are based on models with invariance constraints (see section 5.4.1.3), and estimates of correlated uniquenesses are not displayed for the sake of readability.

Table 90: Fit indices Mastery Experiences simplex models

		Fit indices						*Model comparison**			
	Model	χ^2	*p*	RMSEA	CFI	TLI	SRMR	ΔCFI	$\Delta\chi^2_{(\Delta df)}$	*p*	*pass?*
1	Latent easy level model, strong invariance, simplex	$\chi^2_{(48)}$= 48.08 χ^2/df= 1.00	p=.470	0.00	1.00	1.00	0.08				
1b	Model 1, plus equality constraints on all stability paths	$\chi^2_{(50)}$= 53.00 χ^2/df= 1.06	p=.359	0.02	0.99	0.99	0.13	-0.006	$\Delta\chi^2_{(2)}$= 4.92	p>.05	*yes*
1c	Model 1b, regressed on grade.	$\chi^2_{(61)}$= 73.27 χ^2/df= 1.20	p=.135	0.03	0.98	0.97	0.13				
2	Latent medium level model, *partial* strong invariance, simplex	$\chi^2_{(47)}$= 84.48 χ^2/df= 1.78	p=.001	0.07	0.95	0.93	0.08				
2b	Model 2, plus equality constraints on all stability paths	$\chi^2_{(49)}$= 90.54 χ^2/df= 1.85	p<.001	0.07	0.95	0.93	0.11	-0.005	$\Delta\chi^2_{(2)}$= 6.06	p<.05	*yes*
2c	Model 2b, regressed on grade.	$\chi^2_{(60)}$= 117.10 χ^2/df= 1.95	p<.001	0.07	0.93	0.91	0.11				
3	Latent hard level model, *partial* strong invariance, simplex	$\chi^2_{(47)}$= 73.17 χ^2/df= 1.56	p=.009	0.06	0.97	0.96	0.07				
3b	Model 3, plus equality constraints on all stability paths	$\chi^2_{(49)}$= 76.44 χ^2/df= 1.56	p=.007	0.06	0.97	0.96	0.08	-0.002	$\Delta\chi^2_{(2)}$= 3.27	p>.05	*yes*
3c	Model 3b, regressed on grade.	$\chi^2_{(60)}$= 100.34 χ^2/df= 1.67	p<.001	0.06	0.96	0.94	0.08				

**Models 1b, 2b, and 3b are compared against Models 1, 2, and 3, respectively*

The autoregressive simplex models for mastery experiences indicated that for all three latent constructs, the rank-order of students' mastery experiences were relatively stable across the lessons, with standardised estimates ranging from .74 to .94. In absolute terms, students' mastery experiences for easy and medium difficulty tasks (unstandardised stability paths were .84 and .86, respectively) were more stable than their mastery experiences for hard tasks (unstandardised stability path was .80). Furthermore, grade significantly predicted mastery experiences at the end of lesson 1, where younger students were likely to report higher mastery experience than older students, for all levels of difficulty (standardised estimates ranging from -0.23 for the easy level construct to -0.25 for the hard level construct).



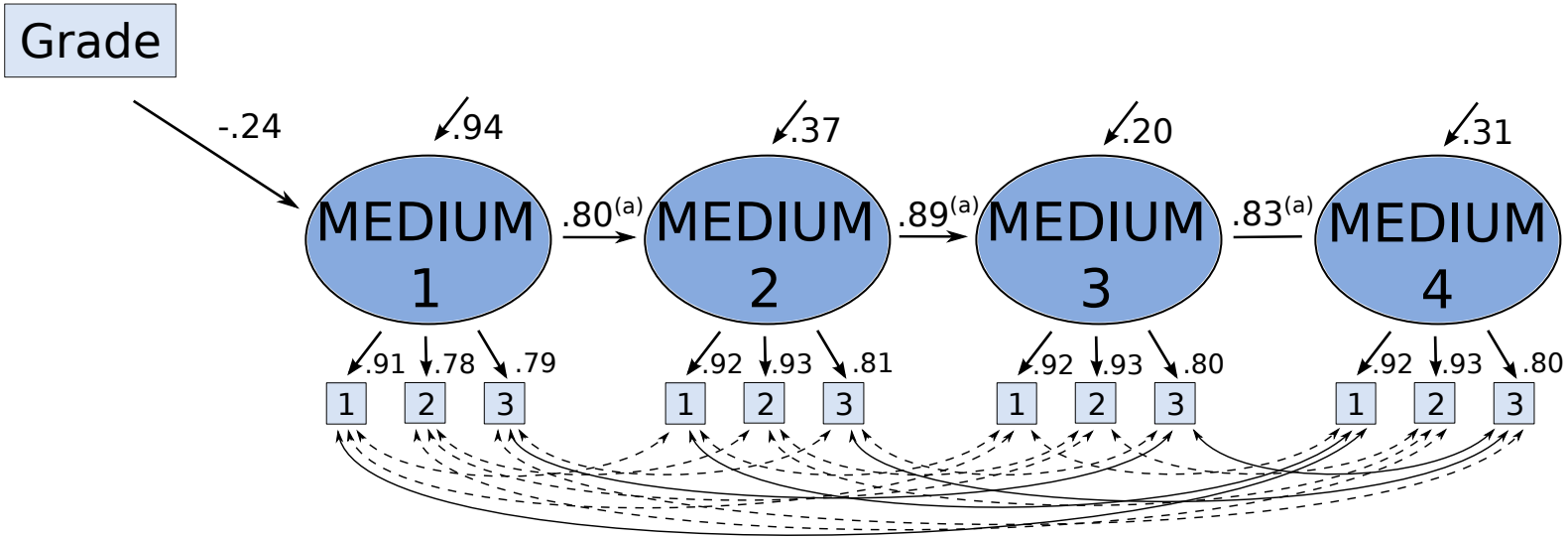
Model fit: $\chi^2_{(61)}$ =73.27, p=.135, RMSEA=.03, CFI=.98, TLI =.97, SRMR=.13

EASY 1 - 4 = Students' Mastery Experiences on easy tasks, lessons 1 - 4

(a) = Equality constraint for stability

Parameter estimates are standardised (STDYX)

Figure 59: Autoregressive model Mastery Experiences, easy tasks



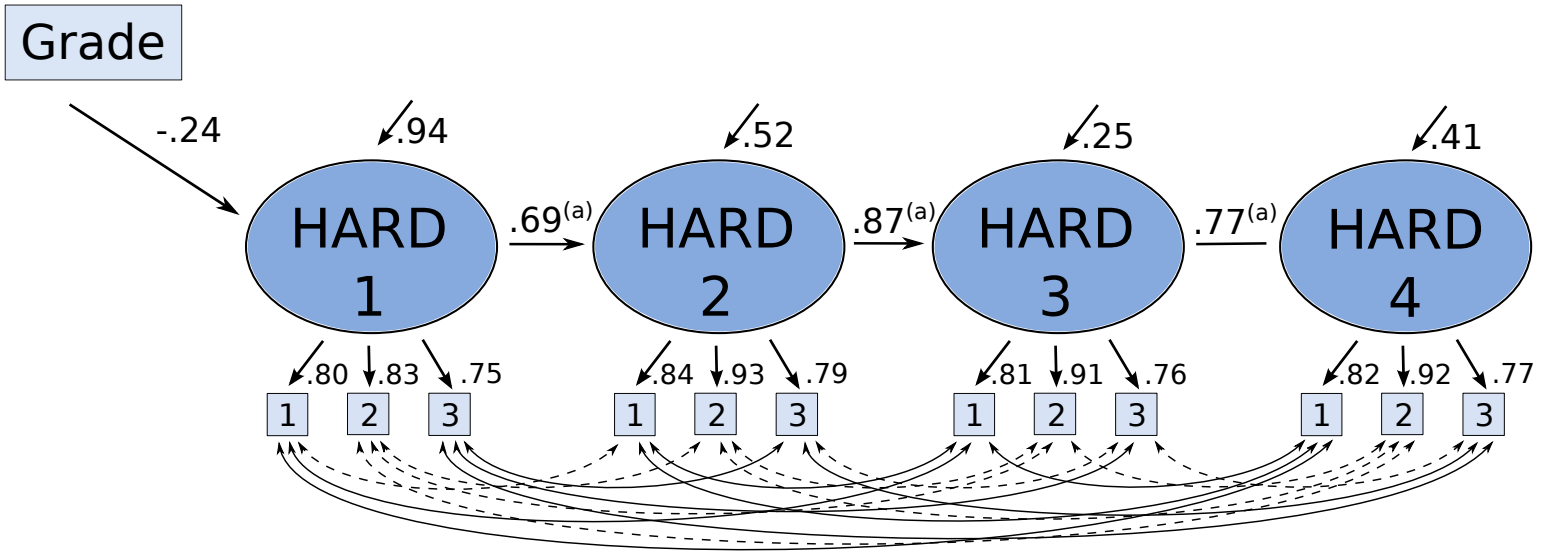
Model fit: $\chi^2_{(60)}$=117.10, p<.001, RMSEA=.07, CFI=.93, TLI =.91, SRMR=.11

MEDIUM 1-4 = Students' Mastery Experiences on medium difficulty tasks, lessons 1-4

(a) = Equality constraint for stability

Parameter estimates are standardised (STDYX)

Figure 60: Autoregressive model Mastery Experiences, medium difficulty tasks



Model fit: $\chi^2_{(60)}$=100.34, p=.001, RMSEA=.06, CFI=.96, TLI =.94, SRMR=.08

HARD 1 - 4 = Students' Mastery Experiences on hard tasks, lessons 1 - 4

(a) = Equality constraint for stability

Parameter estimates are standardised (STDYX)

Figure 61: Autoregressive model Mastery Experiences, hard tasks

5.4.2.2 Stability of the Task self-efficacy measure

Autocorrelations for the Task Self-efficacy measure

Having investigated the stability of the mastery experiences measure, I proceeded to focus on the Task Self-Efficacy measure. In Table 91 I present the autocorrelations for the single Task self-efficacy items, across the four lessons. In this table, the estimates range between $r = .18$ and $r = .59$ (grade 6) and $r = .47$ and $r = .62$ (grade 10). Similar to the multidimensional measures of self-efficacy and mastery experiences, grade 6 correlations are generally of a smaller magnitude than the grade 10 correlation estimates. When inspecting the autocorrelations in regard to the simplex process, there are some higher than expected estimates between the lesson 4 and the lesson 2 autocorrelations, as compared with the lesson 4 and lesson 3 autocorrelations, for both samples. For the grade 6 sample, there is a lower than expected autocorrelation between lessons 1 and 3, as compared with the lesson 1 and 4 estimate. It is worth noting these estimates are on the basis of a single, repeated, item only, which is likely linked with more error and (random) variability (not necessarily true change).

Table 91: Autocorrelations for the single-item Task Self-Efficacy measure

	Item	L1	L2	L3	L4
L1	How certain are you that you can solve similar tasks	1	.59**	.48**	.47**
L2	How certain are you that you can solve similar tasks	.36**	1	.62**	.57**
L3	How certain are you that you can solve similar tasks	.18	.59**	1	.54**
L4	How certain are you that you can solve similar tasks	.36*	.46**	.34*	1

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

Grade 6 estimates are displayed below the diagonal, grade 10 above

Simplex models of Task Self-Efficacy

I specified an autoregressive model for the single Task self-efficacy item across the four lessons to investigate the rank-order stability of students' Task self-efficacy across time (see Table 92). The simplex model resulted in "unacceptable" fit indices according to the a-priori specified thresholds (see section 5.3.3). Based on modification indices as well as my previous inspection of the autocorrelations, I specified an alternative model where students' Task self-efficacy at the end of lesson 4 were predicted in addition by their Task self-efficacy at the end of lesson 2 (Model 2), which resulted in acceptable fit indices. These changes in the model might indicate there was something "common" about lesson 2 and 4, related to students' Task specific self-efficacy.

I proceeded to include equality constraints on the Model 2 stability paths, to investigate whether they differed significantly over time. A model with equality constraints on all adjacent paths (Model 2b) lead to significant detriment in fit. A model with equality constraints on lesson 1-3 paths (Model 2c) passed when considering $\chi 2$ difference, but not when considering CFI. So did a model with equality constraints on adjacent lesson 2-4 paths (Model 2d), which resulted in better fit indices as compared with Model 2c (similar in regard to model parsimony). Considering $\chi 2$ is usually deemed "too sensitive", rather than "not sensitive enough", I consider this as reasonable evidence the paths from lesson 2 to lesson 4 (Model 2d) can be constrained as equal without significant detriment to fit. Overall, this indicates the path from lesson 1 to 2 differed significantly from the following adjacent lesson paths, in particular as compared with the lesson 3 to 4 path.

Table 92: Fit indices Task Self-Efficacy simplex models

		Fit indices						*Model comparison****		
	Model name	χ^2	*p*	RMSEA	CFI	TLI	SRMR	ΔCFI	$\Delta\chi^2_{(\Delta df)}$	*p*	*pass?*
1	Task Self-Efficacy, simplex model	$\chi^2_{(3)}$= 27.09 χ^2/df= 9.03	p<.001	0.23	0.74	0.47	0.11				
2	Task Self-Efficacy, simplex model*	$\chi^2_{(2)}$= 4.86 χ^2/df= 2.43	p= .088	0.10	0.97	0.91	0.05				
2b	Task Self-Efficacy, simplex model*, equality constraints on all adjacent lesson paths	$\chi^2_{(4)}$= 12.71 χ^2/df= 3.18	p= .013	0.12	0.90	0.86	0.14	-0.065	$\Delta\chi^2_{(2)}$= 7.85	p<.05	*no*
2c	Task Self-Efficacy, simplex model*, equality constraints on lesson 1-3 paths	$\chi^2_{(3)}$= 8.09 χ^2/df= 2.70	p= .044	0.11	0.94	0.89	0.10	-0.025	$\Delta\chi^2_{(1)}$= 3.23	p>.05	*yes*
2d	Task Self-Efficacy, simplex model*, equality constraints on lesson 2-4 paths	$\chi^2_{(3)}$= 7.34 χ^2/df= 2.45	p= .062	0.10	0.95	0.91	0.09	-0.017	$\Delta\chi^2_{(1)}$= 2.48	p>.05	*yes*
3	Model 2d, correlated with pre-observations and regressed on grade	$\chi^2_{(150)}$= 269.76 χ^2/df= 1.80	p<.001	0.07	0.94	0.92	0.12				

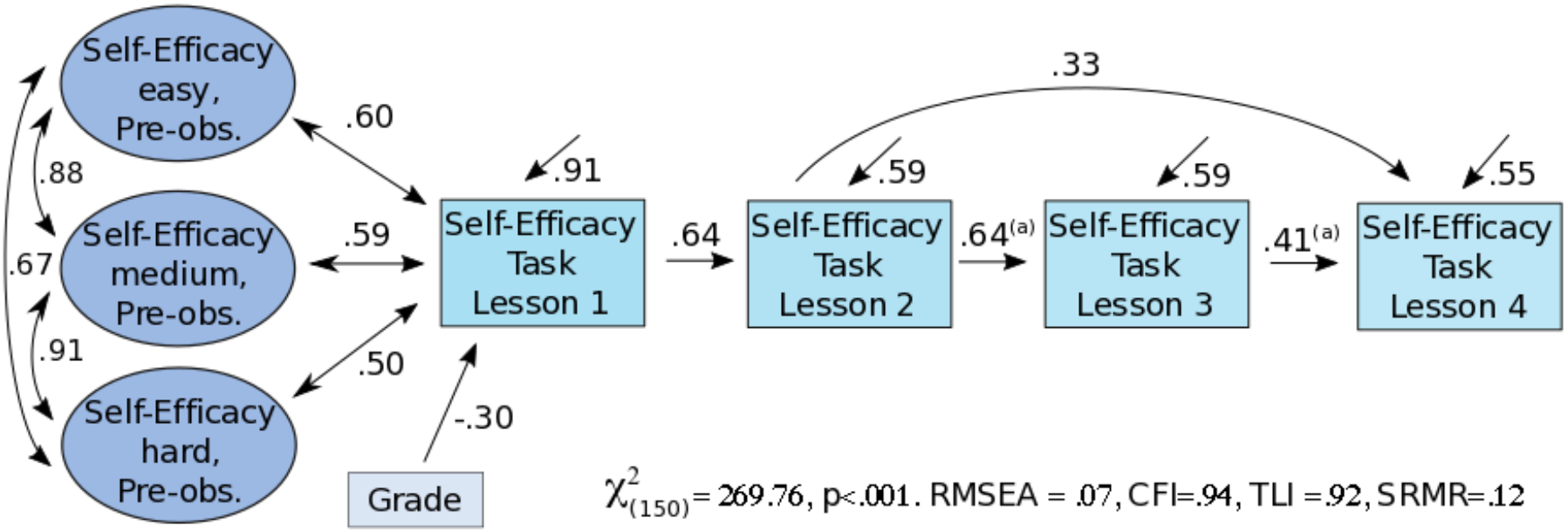
* including stability path from lesson 2 to lesson 4

*** Models 2b-2d were compared against Model 2

On the basis of Model 2d I proceeded to regress lesson 1 Task self-efficacy on student grade and self-efficacy pre-observations. This was in order to account for differences between the grade levels and estimate the relationship between the more general mathematics self-efficacy (at the pre-observation time point) and the more specific Task self-efficacy (at lesson 1). When Task self-efficacy at lesson 1 was predicted by easy, medium, and hard self-efficacy at the pre-observation time points, respectively, all three paths were non-significant. Considering the high correlations between the three pre-observation self-efficacy level constructs, these non-significant estimates are likely related to co-linearity. As such, I tested an equivalent model where these three relationships were specified as correlations rather than unique predictions (Model 3, Table 92), which resulted in identical fit indices. In this case the relationships between the three self-efficacy level constructs at pre-observations and students' Task self-efficacy at lesson 1 were all meaningful and significantly correlated.

Figure 62 displays the parameter estimates from Model 3, including equality constraints on stability parameters from lesson 2 to 3 and 3 to 4. Students' Task self-efficacy appears relatively stable across time, where the rank-order of Task self-efficacy at each lesson predicts the rank-order standing of the subsequent lesson. While the standardised estimate for stability from lesson 1 to lesson 2 appears identical to the lesson 2 to 3 estimate, a model with all adjacent lesson paths constrained as equal led to significant detriment in fit (Model 2b, Table 92), indicating that the lesson 1 to 2 stability path (unstandardised estimate at .80) differed significantly from the lesson 2 to 3 and 3 to 4 stability paths (unstandardised estimates at .50). In addition to the autoregressive paths between adjacent lessons, a path from lesson 2 to lesson 4 was added in order to achieve acceptable model fit. This indicates students' lesson 2 Task self-efficacy scores significantly predicted their rank-order at the end of lesson 4. Students' Task self-efficacy at the end of lesson 1 were strongly related to their pre-

observation self-efficacy for all levels of task difficulty. Furthermore, they were predicted (negatively) by grade, where older students were likely to report weaker Task self-efficacy.



Pre-obs. = Pre-observations, Mastery Exp. = Mastery Experiences
(a) = Equality constraint on stability path. Parameter estimates are standardised (STDYX).
NB! For readability, pre-observation measurement models are not included

Figure 62: Autoregressive path model of Task Self-Efficacy

5.4.2.3 Models including cross-lagged paths between self-efficacy and mastery experiences

I next included measures of self-efficacy (the SEGD short was investigated in study 3, section 4.4.3, and the Task Self-Efficacy measure in section 5.4.4.2) and mastery experiences (investigated in section 5.4.2.1) in cross-lagged path models for each of the levels of difficulty. In order to specify these models, I drew on the previously established factor structures for each of the measures, and created for each of my multidimensional measures latent level parcels (e.g. easy, medium difficulty, and hard parcels). I included equality constraints on the stability (autoregressive) paths according to the previously tested item-level models. I included in the model specifications also grade and students' pre-observation self-efficacy as covariates, predicting each of the other constructs at lesson 1. Fit indices from the cross-lagged path models including easy (Model 1), medium difficulty (Model 2), and hard (Model 3) level constructs, based on the best-fitting autoregressive models with equality constraints for stability parameters, are displayed in Table 93.

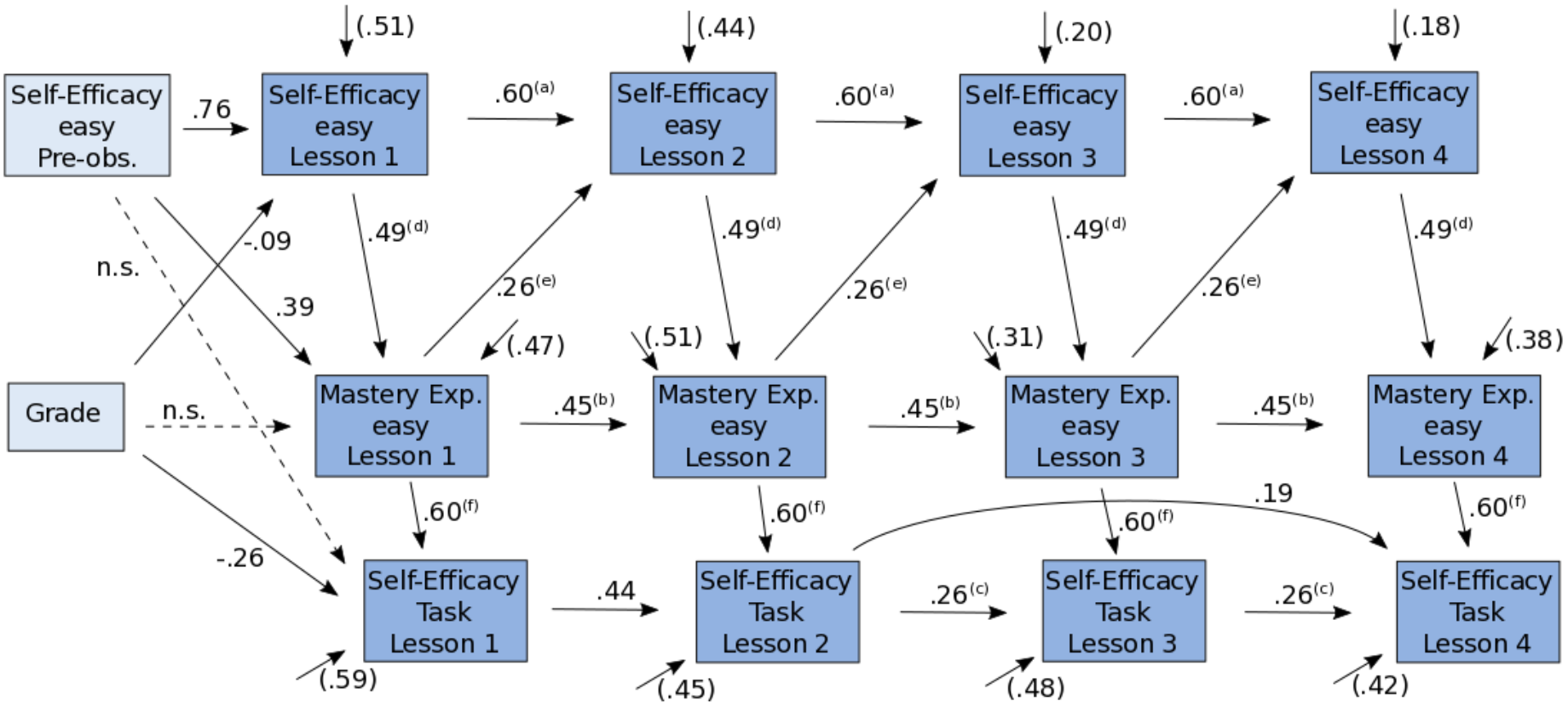
I next included equality constraints also on all cross-lagged paths, in the models for each of the three level constructs. As displayed in Table 93, the models including equality constraints on all cross-lagged parameters (Models 1b, 2b, and 3b, respectively) did not differ significantly to the models without. This indicates the cross-lagged relationships between my three repeated measures did not differ significantly across time in my sample, in relation to either of the level constructs. When considering the fit indices in an absolute sense, the models are below the "acceptable" threshold in the case of easy and medium difficulty constructs (Models 1b and 2b), even though the fit indices do not differ significantly as compared with the less constrained models (e.g. Models 1 and 2). Considering autocorrelations and modification indices, I could easily achieve "acceptable" model fit

through releasing one of the cross-lagged paths. As I discussed in section 2.3.5, study 1, fit indices should not replace sound theoretical judgment. Given I do not have a strong theoretical reason to assume the cross-lagged paths differ in magnitude across time, it seems reasonable to consider model parsimony rather than strict cut-off criteria. As such, I choose to accept and keep models 1b and 2b, although they fail (just) to achieve "acceptable" fit indices. Parameter estimates from Models 1b, 2b, and 3b are displayed in Figures 63, 64, and 65, respectively.

Table 93: Fit indices cross-lagged path models with equality constraints

		Fit indices						*Model comparison****			
	Model name	χ^2	*p*	RMSEA	CFI	TLI	SRMR	ΔCFI	$\Delta\chi^2_{(\Delta df)}$	*p*	*pass?*
1	Autoregressive and cross-lagged paths between easy SE & ME constructs, on grade and preobs. Equality constr. on stability parameters according to established models	$\chi^2_{(68)}$= 153.19 χ^2/df= 2.25	p<.001	0.09	0.91	0.88	0.09				
1b	Model 1 + equality constraints on all cross-lagged parameters	$\chi^2_{(76)}$= 167.33 χ^2/df= 2.20	p<.001	0.09	0.90	0.88	0.13	-0.007	$\Delta\chi^2_{(8)}$= 14.14	p>.05	*yes*
2	Autoregressive and cross-lagged paths between medium SE & ME constructs, on grade and preobs. Equality constr. on stability parameters according to established models	$\chi^2_{(68)}$= 185.24 χ^2/df= 2.72	p<.001	0.10	0.91	0.88	0.08				
2b	Model 2 + equality constraints on all cross-lagged parameters	$\chi^2_{(76)}$= 195.18 χ^2/df= 2.57	p<.001	0.10	0.91	0.89	0.11	-0.002	$\Delta\chi^2_{(8)}$= 9.94	p>.05	*yes*
3	Autoregressive and cross-lagged paths between hard SE & ME constructs, on grade and preobs. Equality constr. on stability parameters according to established models	$\chi^2_{(68)}$= 149.17 χ^2/df= 2.19	p<.001	0.09	0.93	0.91	0.08				
3b	Model 3 + equality constraints on all cross-lagged parameters	$\chi^2_{(76)}$= 156.54 χ^2/df= 2.06	p<.001	0.08	0.93	0.92	0.10	0.001	$\Delta\chi^2_{(8)}$= 7.37	p>.05	*yes*

*** Models 1b, 2b, and 3b are compared against Models 1, 2, and 3, respectively



Pre-obs. = Pre-observations,
Mastery Exp. = Mastery Experiences
Path estimates are unstandardised
Estimates in brackets are standardised (STDYX)
(a), (b), and (c) are equality constraints on stability paths
(d), (e), and (f) are equality constraints on cross-lagged paths

Model fit: $\chi^2_{(75)} = 167.33$, p<.001. CFI=.90, TLI =.88, RMSEA=.09, SRMR=.13
Stability: **a - b** = n.s.; **a - c** = .34, p<.05; **b - c** = n.s.
Direction: **d - e** = .23, p=.06; **d - f** = n.s.; **e - f** = -.34, p<.05

Figure 63: Cross-lagged path model of Self-Efficacy and Mastery Experiences, easy level constructs

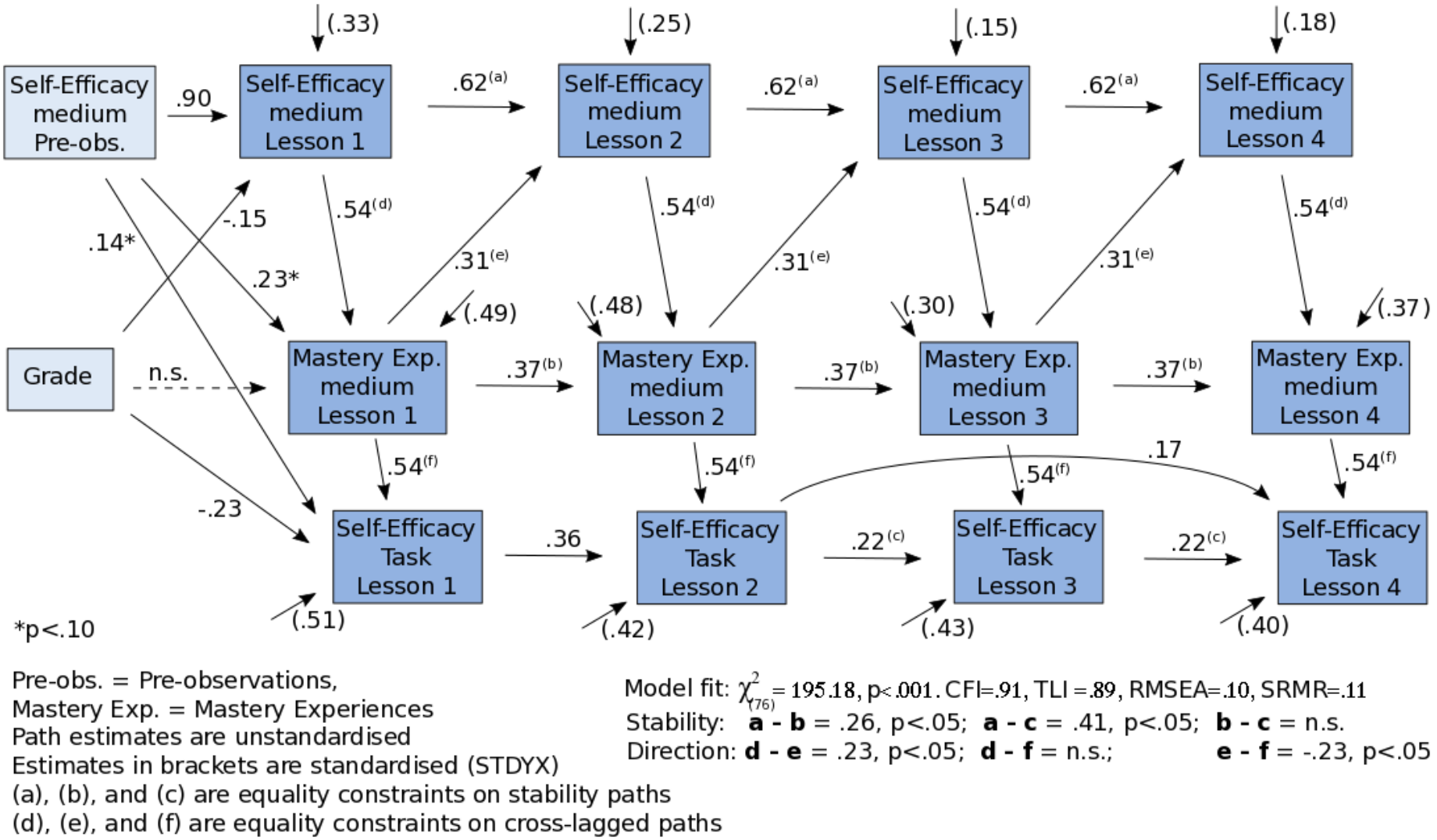


Figure 64: Cross-lagged path model of Self-Efficacy and Mastery Experiences, medium difficulty level constructs

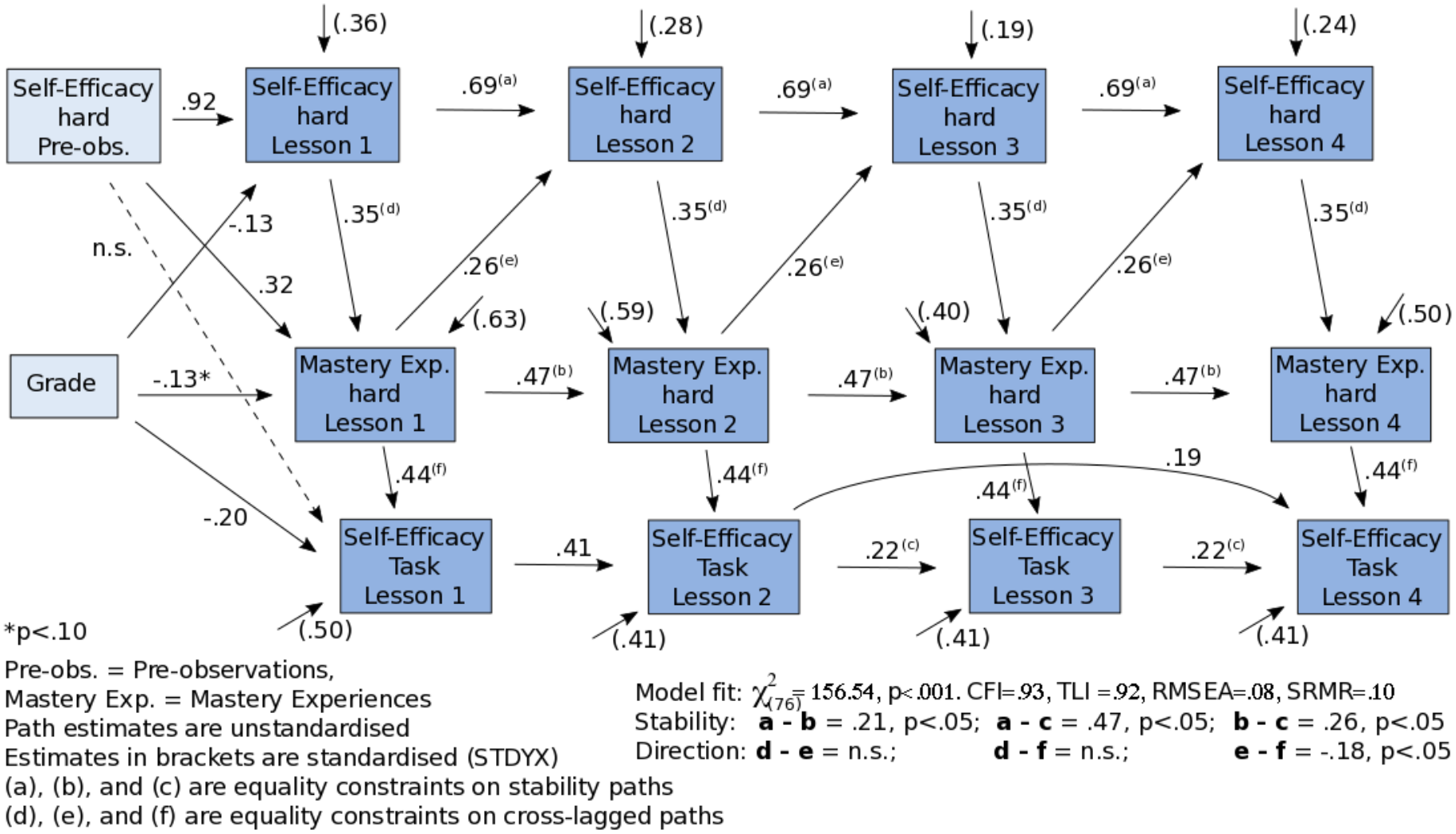


Figure 65: Cross-lagged path model of Self-Efficacy and Mastery Experiences, hard level constructs

The equality constraints enabled statistical comparison of the magnitude of the stability of the latent constructs over time, as well as the cross-lagged paths between them. In regard to stability, students' self-efficacy (as measured at the beginning of each lesson in mathematics) were the most stable (.60, .62, and .69, for easy, medium difficulty, and hard level tasks, respectively), followed by their mastery experiences (.45, .37, and .47, for easy, medium difficulty, and hard level tasks, respectively), and finally their Task self-efficacy (about .40 for the lesson 1-2 path, and about .22 for the lesson 2-3 and 3-4 paths). While in an absolute sense this was true for all levels of difficulty, pairwise comparisons of the stability paths led to statistically significant differences in regard to the hard level constructs (see Figure 65), while not all the differences were statistically significant in regard to the easy and medium difficulty level constructs (see Figures 63 and 64). Looking across the levels of difficulty, self-efficacy was most stable (in an absolute sense) in regard to hard level tasks, followed by medium difficulty and then easy tasks. Students' mastery experiences were comparably stable for easy and hard tasks, and the least stable in regard to medium difficulty level tasks. These differences could not be tested statistically in the present study, as the different levels of difficulty were not in the same model. The Task self-efficacy measure including just a single item was the same in Model 1b, 2b, and 3b, of course, reflected in the small differences in the estimated stability paths across the three models.

Considering the cross-lagged paths between the multidimensional measures of self-efficacy and mastery experiences, all paths were significant, and could be constrained as equal across time. Students' self-efficacy at the beginning of a lesson predicted their mastery experiences at the end of the lesson to a larger degree than their mastery experiences at the end of a lesson predicted their self-efficacy at the beginning of the subsequent lesson. These differences were statistically significant in regard to easy and medium difficulty constructs ($d - e = .23$ for both

constructs), but not hard constructs. In an absolute sense, both these effects (of self-efficacy on mastery and vice versa) were strongest in regard to medium difficulty tasks, as compared with easy and hard tasks. The paths from mastery experiences to Task self-efficacy, both measured at the end of a lesson, were the strongest cross-lagged paths overall, in an absolute sense. They differed statistically from the cross-lagged paths from mastery experiences to lesson self-efficacy in regard to all levels of difficulty ($e - f$ from -.18 to -.34), while the differences were not significant in comparison with the cross-lagged paths from self-efficacy to mastery experiences. Across the levels of difficulty, the paths from mastery experiences to Task self-efficacy were stronger, in an absolute sense, in regard to easy, followed by medium difficulty, and then hard mastery experiences.

As expected, grade predicted self-efficacy negatively, where students in grade 10 were likely to report lower lesson 1 self-efficacy (-.09, -.15, -.13 for easy, medium difficulty, and hard tasks, respectively) and Task self-efficacy (about -.23) as compared with grade 6 students. Grade level also predicted students' lesson 1 mastery experiences negatively, but the effect was only marginally significant, and in the case of hard level tasks (-.13). Furthermore, students' pre-observation self-efficacy predicted their lesson 1 self-efficacy (.76, .90, and .92, for easy, medium difficulty, and hard level tasks), mastery experiences (.39, .23 marginally significant, and .32 for easy, medium difficulty, and hard level tasks), and Task self-efficacy (medium difficulty tasks only, .14, marginally significant).

5.4.2.4 Summary of results for RQ2

In order to investigate the relationships between students' (lesson) self-efficacy, mastery experiences, and Task self-efficacy over time, I included each of the repeated measures in a path model including autoregressive as well as cross-lagged paths. The stability of my multidimensional measures of self-efficacy and mastery experiences did not differ across the

lessons, while the stability of students' Task self-efficacy was higher from lesson 1 to 2, as compared with the subsequent lessons. Overall, students' (lesson) self-efficacy was the most stable, followed by their mastery experiences, and finally their Task self-efficacy. The differences in stability were larger with higher task difficulty, demonstrated in how all differences were statistically significant in the case of hard tasks, two out of three were significant in the case of medium difficulty tasks, while one difference was significant in the case of easy tasks.

I included in each of the three path models (for easy, medium difficulty and hard level tasks) equality constraints on the cross-lagged paths. The fit indices indicated the cross-lagged effects did not differ significantly over time. All cross-lagged and autoregressive paths were significant, supporting a reciprocal effects model of students' self-efficacy and mastery experiences. The effects from students' (lesson) self-efficacy on their mastery experiences were in this study stronger than the effect in the opposite direction, with significant differences in the case of easy and medium difficulty tasks. Furthermore, the effects of students' mastery experiences on their Task self-efficacy were stronger than the effects on their (lesson) self-efficacy at the beginning of the next lesson in mathematics. Considering task difficulty, in my study construct stability was most stable in regard to hard tasks, while cross-lagged effects were weakest in regard to hard tasks, and strongest in regard to medium difficulty tasks. These differences were not tested for statistical significance.

5.4.3 RQ3: The relationship between students' self-efficacy, mastery experiences, and ability conceptions

Given that I adapted an established measure of ability conceptions (Dweck, 2008) for the purposes of my study, I first investigated the factor structure of the adapted ability conceptions measure, before including this measure in the path models developed in research

question 2, including the cross-lagged relationships between self-efficacy and mastery experiences. Investigating the relationships in research question 3 served two purposes: while one aim was to contribute to theory, I also validated my new measures through testing theoretically expected relationships with an established measure. In study 1, the validity of my proposed measure of students' self-efficacy for an upcoming test (the SEGD) was supported from my findings of a positive relationship with students' national test scores (see section 2.5.2). For each of the repeated measures in the present study (e.g. the multidimensional self-efficacy and mastery experiences measures, and the task self-efficacy measure), I included my constructs of ability conceptions as covariates in the previously developed autoregressive models (see sections 4.4.3, 5.4.2.1, and 5.4.2.2, respectively). Following on from these checks of theoretically expected relationships, I proceeded to include my constructs of ability conceptions as covariates in the cross-lagged path models developed for research question 2 (see section 5.4.2.3), in order to investigate direct and indirect effects of ability conceptions on my proposed measures.

5.4.3.1 Factor structure of the ability conceptions measure

To investigate the factor structure of my adapted measure of students' ability conceptions, I first inspected the correlation matrix, before I tested alternative factor structures through confirmatory factor analyses. In Table 94 I present the correlation matrix for the items used in the ability conceptions measure. Note that items 5 to 8 are the original items from Dweck (2008) related to general ability, while items 1 to 4 are adapted for the present purposes, to relate to students' conceptions of ability in mathematics. As expected, entity items (e.g. negatively worded items) are positively correlated with entity items, and incremental items (e.g. positively worded items) are positively correlated with incremental items, while entity items are negatively correlated with incremental items. The strongest correlations appear to be for corresponding items, e.g. correlations between two positively or two negatively worded

items, as compared with correlations between one positively and one negatively worded item. This is indicated through the ranges of estimates, where the correlation estimates for similarly worded items ranges from .25 to .78, while the correlation estimates for items of opposite direction are somewhat weaker, ranging from -.14 to -.53. Furthermore, all five non-significant correlations include oppositely worded items.

Table 94: Correlation matrix Ability Conceptions

	Item	1	2	3	4	5	6	7	8
1	People CANNOT change whether they are good at mathematics or not	1	-.53**	-.37**	.78**	-.17	.37**	-.24*	.41**
2	Even how good one really is at mathematics one can change a lot	-.37**	1	.47**	-.53**	.25*	-.34**	.36**	-.39**
3	No matter who someone is, they can always become better at mathematics if they try	-.14	.52**	1	-.38**	.43**	-.20*	.47**	-.28**
4	To be honest, people CANNOT change how good they are at mathematics	.69**	-.38**	-.33**	1	-.23*	.42**	-.19	.43**
5	No matter who someone is, they can significantly change how smart they are	-.44**	.40**	.40**	-.46**	1	-.49**	.61**	-.42**
6	How smart someone is, is something they CANNOT change very much	.50**	-.42**	-.24*	.56**	-.38**	1	-.50**	.57**
7	People can always change how smart they are a lot	-.21	.33**	.26*	-.27*	.49**	-.28*	1	-.52**
8	People can learn new things, but they CANNOT really change how smart they are	.32**	-.29*	-.30*	.41**	-.30*	.53**	-.15	1

** Correlation is significant at the 0.01 level (2-tailed), * correlation is significant at the 0.05 level (2-tailed).
Grade 6 estimates are displayed below the diagonal, grade 10 above

I tested several theoretically plausible, alternative factor structures. As a baseline or comparison model, I specified a Unidimensional model (Model 0), where all items loaded on a single ability conceptions factor. Previous studies (e.g. Bonne & Johnston, 2016) have posited two factors of ability conceptions (entity and incremental beliefs), which I specified in Model 1. In contrast to this conceptualisation, Marsh et al. (2010) suggested that negatively worded items might best be captured by a method factor (or correlated uniquenesses), as opposed to constituting a substantive factor. I tested this approach in Models 0b and 0c, in which correlated uniquenesses or a latent method factor are included in the specification, together with the single ability conceptions factor. In my adaptation of the original measure I included also a dimension in regard to the specificity of the items: items related to general ability versus items related to ability in mathematics. I tested this alternative factor structure in Model 2, in which I specified two latent factors (general ability conceptions and mathematics ability conceptions), together with correlated uniquenesses accounting for the negatively worded items. In line with model 0c, I specified also a model in which the negatively worded items loaded on a latent method factor (Model 2b). Finally, on the basis of the factor structures of Models 1 and 2, I tested two higher-order models (Models 1b and 2c).

Table 95: Results confirmatory factor analysis, Ability Conceptions

		Fit indices					
	Model	χ^2	*p*	RMSEA	CFI	TLI	SRMR
0	Unidimensional model	$\chi^2_{(20)}$= 109.13 χ^2/df= 5.46	*p*<.001	0.16	0.72	0.60	0.09
0b	Unidimensional model, CUs for negative wording	$\chi^2_{(14)}$= 42.96 χ^2/df= 3.07	*p*<.001	0.11	0.91	0.82	0.06
0c	Unidimensional model, method factor for negative wording	$\chi^2_{(15)}$= 53.18 χ^2/df= 3.54	*p*<.001	0.12	0.88	0.77	0.06
1	2 latent constructs (entity & incremental)	$\chi^2_{(19)}$= 76.18 χ^2/df= 4.00	*p*<.001	0.13	0.82	0.73	0.08
1b	Second order model: ability BY entity & incremental	*Did not converge*					
2	2 latent constructs (general & mathematics ability), CUs for neg. wording	$\chi^2_{(13)}$= 18.53 χ^2/df= 1.43	*p*=.139	0.05	0.98	0.96	0.04
2b	2 latent constructs (general & mathematics ability), method factor for negative wording	$\chi^2_{(15)}$= 24.90 χ^2/df= 1.66	*p*=.051	0.06	0.97	0.94	0.04
2c	Second order model: ability BY math & general. CUs for neg. wording*	$\chi^2_{(12)}$= 17.10 χ^2/df= 1.43	*p*=.146	0.05	0.98	0.96	0.04

*negative residual variance on one factor

According to Table 95 Models 0b, 2, and 2b are "acceptable". Considering the a priori determined χ2 and CFI cut-offs (see section 5.3.3) a two-factor model (Model 2), including latent factors for students' conceptions of mathematics ability and general ability, with correlated uniquenesses for negatively worded items, fit my data best. This model is also substantively meaningful, and results in no out-of-bounds estimates. Accordingly, Model 2 is the measurement model I will apply in the proceeding analyses. For the correlation matrices the negatively worded items were reversed so that for all items, higher scores equal stronger beliefs in fixed ability. Accordingly, I will from now on use the term "fixed beliefs" (in general ability and in mathematics ability, respectively) in order to indicate the direction of the scoring, and in order not to confuse my constructs with the entity and incremental constructs from Model 1 (see Table 95) and previous studies.

5.4.3.2 Correlation matrix including latent self-efficacy, mastery experiences and ability conceptions factors

Based on the factor structures of my three latent measures (students' self-efficacy, mastery experiences, and ability conceptions) I included 29 latent factors (including the latent self-efficacy and mastery experiences factors repeated over time) and four Task self-efficacy items in a correlation matrix, in order to inspect the cross-correlation estimates. As can be seen in Table 96 in association with factors 1 and 2, the grade 6 correlations are generally higher than the grade 10 correlations, indicating a stronger (negative) relationship between younger, as compared with older, students' fixed ability beliefs and their self-efficacy and mastery experiences. Furthermore, there is a tendency that students' fixed beliefs in mathematics ability are more strongly correlated than their fixed beliefs in general ability with their self-efficacy and mastery experiences, respectively. This is demonstrated in the generally larger magnitude for cross-correlations with mathematics ability fixed beliefs as compared with general ability fixed beliefs, and in the number of significant cross-correlations (grade 6: 24 of 31 mathematics ability cross-correlations are significant versus 19 of 31 for general ability, grade 10: 18 of 31 mathematics ability cross-correlations are significant versus only 3 of 31 for general ability). This indicates there are substantive differences between the two fixed beliefs constructs, in support of the measurement model derived in section 5.4.3.1.

Table 96: Correlation matrix Ability Conceptions, Self-efficacy, and Mastery Experiences, lessons 1-4

	Factors	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
1	Fixed beliefs, mathematics ability	1	.49**	-.35**	-.30**	-.15	-.31**	-.29**	-.23*	-.22	-.16	-.01	-.34**	-.24*	-.08	-.37**	-.27*	-.09	-.42**	-.31**	-.21	-.35**	-.25*	-.14	-.23*	-.17	-.09	-.31*	-.31*	-.26*	-.19	-.20	-.05	-.34**
2	Fixed beliefs, general ability	.66**	1	-.18	-.11	.06	-.19	-.12	-.01	-.23*	-.16	.05	-.24*	-.12	.03	-.28*	-.22	-.07	-.18	-.06	.04	-.15	-.08	.00	-.18	-.17	-.11	-.22	-.15	-.16	.08	-.01	.07	-.13
3	Pre-obs Self-Efficacy, easy	-.48**	-.30*	1	.86**	.67**	.73**	.74**	.63**	.75**	.77**	.56**	.77**	.78**	.62**	.77**	.72**	.51**	.65**	.69**	.62**	.77**	.78**	.73**	.62**	.60**	.60**	.73**	.72**	.60**	.54**	.67**	.52**	.58**
4	Pre-obs Self-Efficacy, medium	-.46**	-.29*	.74**	1	.86**	.69**	.87**	.81**	.63**	.82**	.70**	.66**	.82**	.74**	.68**	.79**	.65**	.57**	.65**	.62**	.61**	.69**	.71**	.55**	.61**	.65**	.68**	.72**	.66**	.55**	.65**	.54**	.61**
5	Pre-obs Self-Efficacy, hard	-.32**	-.27*	.43**	.82**	1	.50**	.75**	.81**	.45**	.71**	.74**	.45**	.68**	.76**	.45**	.62**	.63**	.42**	.50**	.51**	.49**	.56**	.67**	.39**	.46**	.55**	.51**	.56**	.55**	.51**	.59**	.48**	.47**
6	L1 Self-Efficacy, easy	-.49**	-.32**	.66**	.51**	.32**	1	.88**	.67**	.69**	.72**	.54**	.80**	.80**	.61**	.83**	.82**	.62**	.76**	.79**	.69**	.50**	.59**	.58**	.57**	.64**	.64**	.68**	.66**	.57**	.51**	.45**	.54**	.51**
7	L1 Self-Efficacy, medium	-.41**	-.33**	.54**	.70**	.60**	.76**	1	.88**	.66**	.81**	.70**	.69**	.83**	.72**	.75**	.85**	.74**	.66**	.72**	.68**	.48**	.59**	.63**	.51**	.62**	.68**	.64**	.67**	.62**	.57**	.53**	.57**	.54**
8	L1 Self-Efficacy, hard	-.33**	-.26*	.34**	.64**	.75**	.53**	.78**	1	.51**	.73**	.82**	.52**	.72**	.78**	.54**	.71**	.80**	.59**	.64**	.62**	.43**	.55**	.61**	.44**	.56**	.66**	.55**	.60**	.60**	.54**	.50**	.54**	.51**
9	L2 Self-Efficacy, easy	-.45**	-.26	.55**	.46**	.26*	.74**	.65**	.43**	1	.89**	.59**	.91**	.81**	.55**	.91**	.83**	.52**	.73**	.69**	.56**	.72**	.70**	.59**	.81**	.74**	.60**	.69**	.63**	.56**	.53**	.51**	.51**	.44**
10	L2 Self-Efficacy, medium	-.41**	-.28*	.48**	.61**	.53**	.62**	.87**	.70**	.72**	1	.82**	.84**	.92**	.77**	.82**	.91**	.72**	.70**	.73**	.66**	.66**	.71**	.65**	.72**	.74**	.68**	.67**	.68**	.63**	.62**	.55**	.53**	.51**
11	L2 Self-Efficacy, hard	-.20	-.13	.22	.48**	.58**	.38**	.69**	.82**	.49**	.85**	1	.55**	.77**	.86**	.49**	.71**	.83**	.53**	.62**	.65**	.45**	.55**	.58**	.46**	.53**	.56**	.47**	.56**	.57**	.56**	.44**	.48**	.46**
12	L3 Self-Efficacy, easy	-.28*	-.16	.44**	.38**	.22	.66**	.56**	.40**	.70**	.60**	.42**	1	.88**	.64**	.92**	.82**	.55**	.78**	.74**	.62**	.79**	.77**	.67**	.76**	.75**	.66**	.74**	.69**	.60**	.40**	.56**	.56**	.58**
13	L3 Self-Efficacy, medium	-.34**	-.30*	.59**	.57**	.38**	.71**	.74**	.57**	.60**	.80**	.63**	.75**	1	.87**	.84**	.93**	.78**	.74**	.79**	.74**	.74**	.82**	.82**	.76**	.82**	.78**	.75**	.78**	.75**	.57**	.69**	.64**	.65**
14	L3 Self-Efficacy, hard	-.24	-.22	.33**	.48**	.50**	.46**	.59**	.75**	.37**	.70**	.84**	.41**	.71**	1	.59**	.78**	.85**	.58**	.67**	.69**	.58**	.70**	.78**	.59**	.70**	.74**	.62**	.69**	.70**	.50**	.61**	.62**	.62**
15	L4 Self-Efficacy, easy	-.34*	-.25	.33*	.41**	.36*	.37**	.36**	.36**	.74**	.63**	.56**	.60**	.54**	.33*	1	.90**	.63**	.84**	.78**	.63**	.63**	.63**	.54**	.78**	.75**	.67**	.80**	.72**	.62**	.49**	.40**	.60**	.61**
16	L4 Self-Efficacy, medium	-.36*	-.42**	.15	.59**	.66**	.21	.64**	.63**	.42**	.78**	.74**	.45**	.60**	.57**	.63**	1	.85**	.79**	.83**	.74**	.61**	.69**	.66**	.76**	.82**	.79**	.80**	.80**	.73**	.63**	.52**	.72**	.69**
17	L4 Self-Efficacy, hard	-.35*	-.42**	.19	.57**	.71**	.19	.48**	.77**	.22	.67**	.80**	.25	.57**	.86**	.48**	.76**	1	.64**	.72**	.72**	.43**	.55**	.61**	.57**	.67**	.72**	.67**	.69**	.67**	.59**	.44**	.69**	.63**
18	L1 Mastery Experiences, easy	-.51**	-.29*	.56**	.54**	.29*	.61**	.67**	.38**	.59**	.74**	.50**	.38**	.60**	.38**	.45**	.49**	.25	1	.93**	.77**	.70**	.71**	.60**	.70**	.70**	.63**	.74**	.68**	.58**	.61**	.59**	.54**	.54**
19	L1 Mastery Experiences, medium	-.38**	-.18	.44**	.55**	.39**	.48**	.70**	.50**	.53**	.77**	.65**	.42**	.53**	.48**	.42**	.69**	.45**	.84**	1	.91**	.68**	.76**	.71**	.65**	.68**	.65**	.73**	.74**	.70**	.70**	.65**	.59**	.57**
20	L1 Mastery Experiences, hard	-.26*	-.09	.32*	.57**	.49**	.32**	.59**	.58**	.36**	.61**	.61**	.25	.41**	.49**	.36**	.56**	.55**	.61**	.86**	1	.57**	.68**	.68**	.53**	.57**	.58**	.65**	.68**	.66**	.75**	.61**	.50**	.55**
21	L2 Mastery Experiences, easy	-.45**	-.34**	.60**	.55**	.40**	.71**	.58**	.41**	.75**	.62**	.38**	.71**	.67**	.41**	.65**	.35*	.33*	.53**	.46**	.40**	1	.95**	.80**	.81**	.70**	.55**	.79**	.80**	.70**	.53**	.77**	.51**	.59**
22	L2 Mastery Experiences, medium	-.36**	-.32**	.54**	.62**	.48**	.61**	.63**	.51**	.64**	.61**	.44**	.56**	.70**	.49**	.70**	.50**	.46**	.48**	.43**	.39**	.85**	1	.90**	.76**	.74**	.66**	.81**	.87**	.79**	.60**	.84**	.57**	.62**
23	L2 Mastery Experiences, hard	-.18	-.21	.32**	.54**	.53**	.29*	.50**	.52**	.40**	.52**	.47**	.37**	.59**	.53**	.47**	.47**	.50**	.31*	.41**	.47**	.59**	.81**	1	.62**	.70**	.74**	.69**	.81**	.80**	.54**	.83**	.61**	.57**
24	L3 Mastery Experiences, easy	-.30*	-.11	.39**	.37**	.29*	.31*	.33*	.34**	.57**	.52**	.47**	.46**	.55**	.38**	.85**	.46**	.43**	.44**	.44**	.43**	.66**	.67**	.52**	1	.92**	.77**	.76**	.69**	.64**	.38**	.51**	.66**	.56**
25	L3 Mastery Experiences, medium	-.48**	-.44**	.47**	.50**	.37**	.36**	.38**	.32*	.56**	.56**	.43**	.44**	.59**	.37**	.77**	.49**	.48**	.50**	.45**	.37**	.73**	.82**	.67**	.87**	1	.93**	.71**	.69**	.67**	.38**	.50**	.72**	.51**
26	L3 Mastery Experiences, hard	-.41**	-.40**	.32*	.42**	.36**	.30*	.25	.36**	.38**	.39**	.34*	.15	.47**	.47**	.43**	.21	.53**	.28*	.27*	.36**	.62**	.69**	.75**	.55**	.73**	1	.66**	.67**	.65**	.37**	.47**	.73**	.46**
27	L4 Mastery Experiences, easy	-.34*	-.25	.23	.31*	.29*	.43**	.46**	.35**	.56**	.57**	.41**	.69**	.51**	.20	.55**	.37**	.20	.31*	.34*	.32*	.65**	.50**	.36**	.48**	.48**	.32*	1	.94**	.80**	.43**	.55**	.53**	.80**
28	L4 Mastery Experiences, medium	-.41**	-.37**	.28*	.44**	.40**	.35**	.41**	.36**	.60**	.58**	.41**	.60**	.52**	.21	.64**	.49**	.37**	.34*	.29*	.21	.73**	.66**	.49**	.50**	.65**	.52**	.74**	1	.92**	.49**	.59**	.51**	.83**
29	L4 Mastery Experiences, hard	-.23	-.31*	.22	.43**	.44**	.13	.28*	.40**	.42**	.47**	.43**	.36**	.37**	.25	.45**	.38**	.39**	.18	.14	.21	.46**	.55**	.56**	.24	.45**	.50**	.54**	.79**	1	.49**	.56**	.48**	.80**
30	L1 Task Self-Efficacy	-.13	.17	.27*	.36**	.25	.34**	.33**	.40**	.39**	.35**	.45**	.21	.22	.22	.43**	.10	.14	.32**	.35**	.39**	.32*	.44**	.33*	.38**	.33*	.20	.15	.15	.16	1	.36**	.18	.36*
31	L2 Task Self-Efficacy	-.19	-.25*	.47**	.45**	.29*	.35**	.34**	.27*	.35**	.41**	.32*	.21	.48**	.40**	.31*	.11	.24	.46**	.35**	.35**	.52**	.56**	.54**	.21	.52**	.60**	.26*	.42**	.49**	.36**	1	.59**	.46**
32	L3 Task Self-Efficacy	-.33*	-.31*	.29*	.28*	.18	.19	.20	.19	.29*	.29*	.23	.43**	.40**	.24	.09	.22	.27	.11	.12	.07	.42**	.41**	.52**	.04	.37**	.53**	.43**	.68**	.65**	.18	.59**	1	.34*
33	L4 Task Self-Efficacy	.04	.00	.04	.18	.24	.00	.13	.20	.09	.36**	.46**	.11	.26	.28*	.19	.23	.30*	.10	.08	.10	.11	.18	.36**	.13	.31*	.32*	.32*	.18	.43**	.36*	.46**	.34*	1

** Correlation is significant at the 0.01 level (2-tailed), * Correlation is significant at the 0.05 level (2-tailed).

L1 - L4 = Lesson 1 - Lesson 4

Grade 6 estimates are displayed below the diagonal, grade 10 above the diagonal

In the case of grade 10 fixed beliefs in mathematics ability, correlations with both self-efficacy and mastery experiences constructs are generally greater in magnitude in regard to the easy level, followed by the medium difficulty and the hard level constructs (average estimate of fixed beliefs in mathematics with self-efficacy easy constructs = -.40; medium constructs = -.29; hard constructs = -.10; average estimates of fixed beliefs in mathematics with mean mastery experiences easy constructs = -.20; medium constructs = -.08; hard constructs = .02). In the case of grade 6 the strength of the estimates are less consistent, where they are some times greater in regard to medium difficulty level constructs (average estimates of fixed beliefs in mathematics with self-efficacy easy constructs = -.47; medium constructs = -.58; hard constructs = -.37; average estimates of fixed beliefs in mathematics with mastery experiences easy constructs = -.65; medium constructs = -.65; hard constructs = -.50).

Similarly, for fixed beliefs in general ability, the grade 10 estimates are generally greater in magnitude in regard to the easy level (self-efficacy = -.31; mastery experiences = -.21), followed by the medium difficulty (self-efficacy = -.16; mastery experiences = -.04) and the hard level constructs (self-efficacy = -.05; mastery experiences = .08). The grade 6 estimates for fixed beliefs in general ability are generally greatest in regard to the medium difficulty level constructs (self-efficacy = -.47; mastery experiences = -.51), followed by hard level constructs (self-efficacy = -.32; mastery experiences = -.46) and easy level constructs (self-efficacy = -.18; mastery experiences = -.41). In the case of the grade 6 students, the correlation estimates between fixed beliefs in mathematics ability (as measured before the first lesson) and self-efficacy are somewhat weaker at the last, as compared with the first lesson, while the correlations between fixed beliefs in general ability and self-efficacy are somewhat stronger at the last, as compared with the first lesson. This may be a function of variability of students' fixed beliefs – I measured students' beliefs at a single time point, and

do not know to what degree they varied over time – or it may be an indication students' drew primarily on their mathematics related ability conceptions when they first engaged with a new mathematical topic, and then drew increasingly also on their general ability conceptions as they progressed in their learning.

The grade 6 cross-correlations in relation to the Task item in Table 96 are inconsistent, in that several of the correlation estimates, including some cross-correlations with self-efficacy and mastery experiences for the same lesson (e.g. lessons 3 and 4), are non-significant. This might indicate the variability of grade 6 students' Task self-efficacy, but it may also reflect the disadvantages of using just a single item as an indicator for a construct, especially with younger students.

5.4.3.3 The relationship between ability conceptions and self-efficacy

I tested first the relationship between ability conceptions and my measures of lesson self-efficacy, e.g. the SEGD, short version. I specified models for students' self-efficacy based on the autoregressive models developed in section 4.4.3.1, study 3, including strong measurement invariance (partial strong measurement invariance in the case of the easy level model) and equality constraints on the stability paths. I regressed each of these models on students' fixed beliefs in mathematics ability and general ability, as per the model structure developed in section 5.4.3.1. These models demonstrated a high correlation (r = .70) between the two fixed beliefs constructs, and contrary to my expectations the relationship with self-efficacy was significant only in the case of fixed beliefs of mathematics, not in relation to fixed beliefs in general ability. Due to the possible issue with co-linearity, I re-specified the relationship with ability conceptions as correlations, rather than unique predictions, which resulted in more meaningful parameter estimates. Table 97 displays the fit indices for each of the models including easy, medium difficulty, and hard self-efficacy constructs, correlated

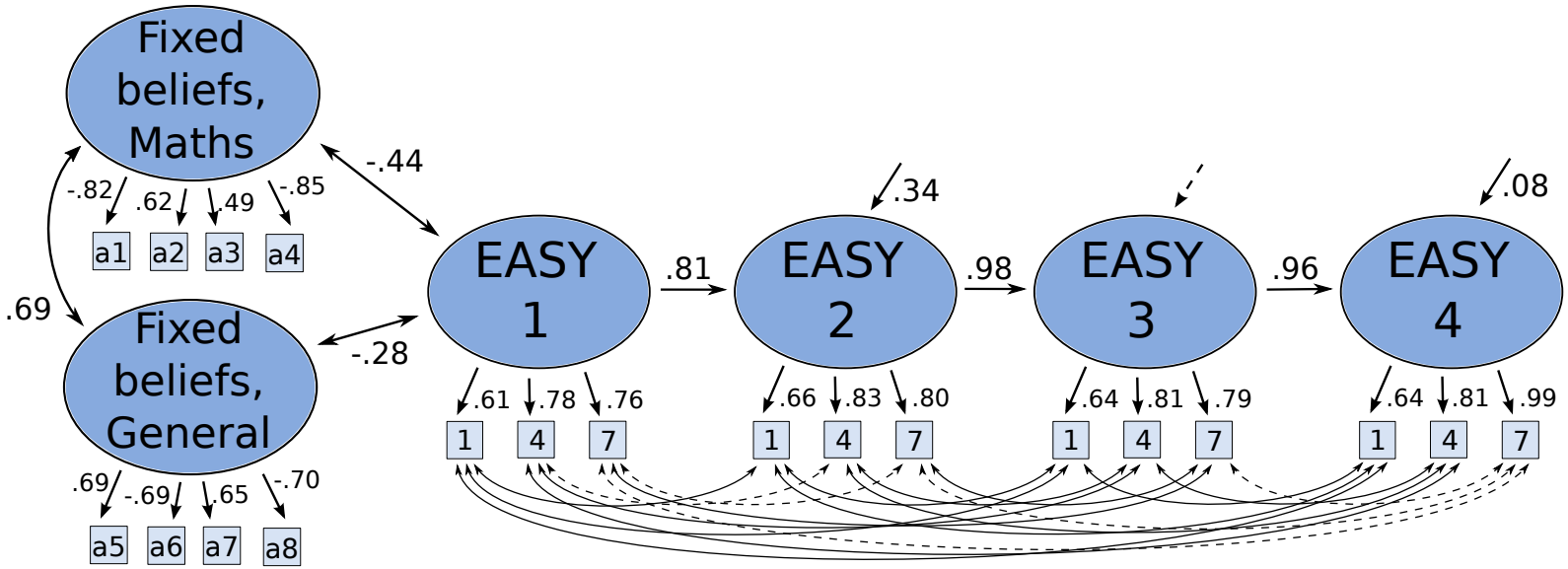
with students' fixed beliefs in mathematics ability and general ability. Model 1 resulted in a negative residual on a single item, where the series of tests recommended by Chen et al. (2001, see section 2.3.5) indicated it was reasonable to fix the residual of this item to a small, positive value (Model 1b). The final model for each of the level constructs (Models 1b, 2, and 3) resulted in acceptable fit indices and meaningful parameter estimates, which are displayed in Figures 66 to 68.

Table 97: Results Self-Efficacy models with Ability Conceptions as covariates

		Fit indices					
	Model	χ^2	*p*	RMSEA	CFI	TLI	SRMR
1	Latent easy level model, *partial* strong invariance, simplex, equality constraints on all paths. Lesson 1 self-efficacy WITH ability conceptions*	$\chi^2_{(162)}$= 292.98 χ^2/df= 1.81	p<.001	0.07	0.90	0.89	0.10
1b	Latent easy level model, *partial* strong invariance, simplex, equality constraints on all paths. Lesson 1 self-efficacy WITH ability conceptions**	$\chi^2_{(163)}$= 292.80 χ^2/df= 1.80	p<.001	0.07	0.90	0.89	0.10
2	Latent medium level model, strong invariance, simplex, equality constraints on all paths. Lesson 1 self-efficacy WITH ability conceptions	$\chi^2_{(163)}$= 278.94 χ^2/df= 1.71	p<.001	0.06	0.93	0.92	0.08
3	Latent hard level model, strong invariance, simplex, equality constraints on all paths. Lesson 1 self-efficacy WITH ability conceptions	$\chi^2_{(163)}$= 268.03 χ^2/df= 1.64	p<.001	0.06	0.94	0.93	0.06

*negative residual variance of one item

** variance of item 3, lesson 4, set at .05



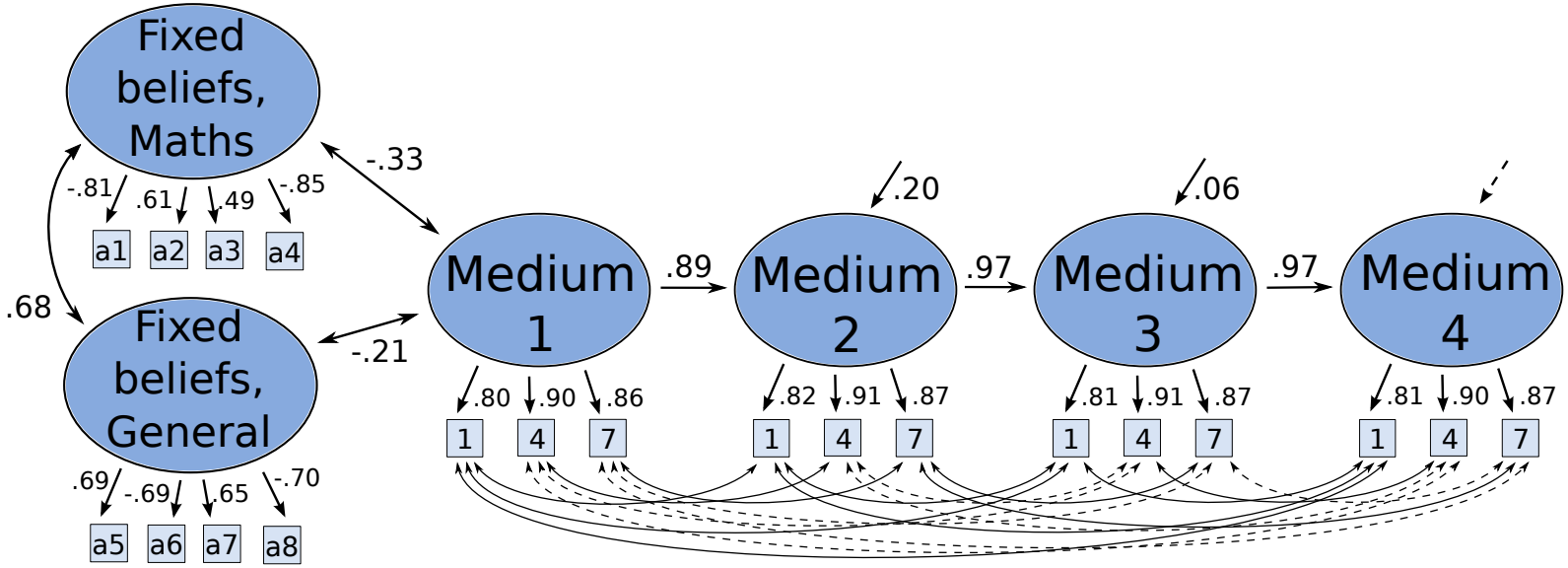
Model fit: $\chi^2_{(163)}$=292.80, p<.001, RMSEA=.07, CFI=.90, TLI =.89, SRMR=.10

Fixed beliefs, Maths / General = Fixed beliefs in mathematics ability / general ability

EASY 1 - 4 = Self-efficacy for easy tasks at the beginning of lessons 1 - 4

Parameter estimates are standardised (STDYX)

Figure 66: Simplex model of Self-Efficacy, easy tasks, with Ability Conceptions



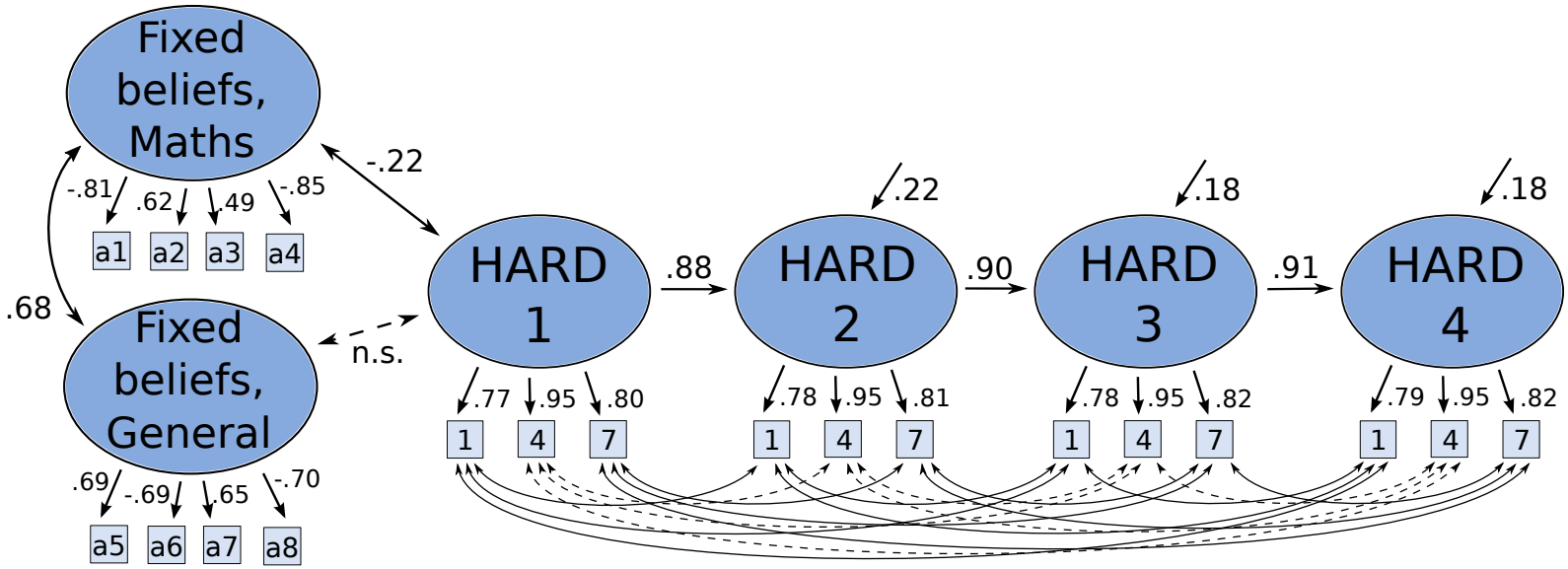
Model fit: $\chi^2_{(163)}$=278.94, p<.001, RMSEA=.06, CFI=.93, TLI =.92, SRMR=.08

Fixed beliefs, Maths / General = Fixed beliefs in mathematics ability / general ability

Medium 1 - 4 = Self-efficacy for medium difficulty tasks at the beginning of lessons 1 - 4

Parameter estimates are standardised (STDYX)

Figure 67: Simples model of Self-Efficacy, medium difficulty tasks, with Ability Conceptions



Model fit: $\chi^2_{(163)}$=268.03, p<.001, RMSEA=.06, CFI=.94, TLI =.93, SRMR=.06

Fixed beliefs, Maths / General = Fixed beliefs in mathematics ability / general ability

HARD 1 - 4 = Self-efficacy for hard tasks at the beginning of lessons 1 - 4

Parameter estimates are standardised (STDYX)

Figure 68: Simplex model of Self-Efficacy, hard tasks, with Ability Conceptions

As can be seen in Figures 66 to 68, students' fixed beliefs in mathematics ability were significantly correlated with their self-efficacy in the case of all three levels of task difficulty, while students' fixed beliefs in general ability were significantly correlated with self-efficacy for easy and medium difficulty tasks. In absolute terms, the strength of the relationship between self-efficacy and ability conceptions was stronger for fixed beliefs in mathematics as compared with fixed beliefs in general ability, and the relationship was progressively weaker with increasing (perceived) task difficulty – however, these tendencies were not tested for statistical significance. Importantly, the significant relationship between self-efficacy and students' fixed beliefs, as measured with items adapted from the established instrument by Dweck (2008), supported the validity of my proposed self-efficacy constructs.

5.4.3.4 The relationship between ability conceptions and mastery experiences

Similar to the models including self-efficacy, I proceeded to specify students' ability conceptions as correlates of their lesson-1 mastery experiences, on the basis of the autoregressive models developed in section 5.4.2.1. These models include partial strong invariance in the case of the easy and the medium difficulty level constructs, and strong invariance in the case of the hard level constructs. Furthermore, for all three level constructs they include equality constraints for the mastery experiences stability paths. As displayed in Table 98, all three models (Models 1-3) resulted in acceptable fit indices, although for medium difficulty and hard tasks, the estimates of CFI are markedly worse as compared to the models including the self-efficacy constructs (e.g. Table 97). The parameter estimates are displayed in Figures 69 to 71.

Table 98: Results Mastery Experiences models w/ Ability Conceptions as covariates

		Fit indices					
	Model	χ^2	*p*	RMSEA	CFI	TLI	SRMR
1	Latent easy level model, strong invariance, simplex, equality constraints on all paths. Lesson 1 mastery WITH ability conceptions	$\chi^2_{(163)}$= 265.40 χ^2/df= 1.63	p<.001	0.06	0.92	0.90	0.11
2	Latent medium level model, *partial* strong invariance, simplex, equality constraints on all paths. Lesson 1 mastery WITH ability conceptions	$\chi^2_{(162)}$= 320.33 χ^2/df= 1.98	p<.001	0.07	0.89	0.87	0.10
3	Latent hard level model, *partial* strong invariance, simplex, equality constraints on all paths. Lesson 1 mastery WITH ability conceptions	$\chi^2_{(162)}$= 317.03 χ^2/df= 1.96	p<.001	0.07	0.89	0.87	0.10

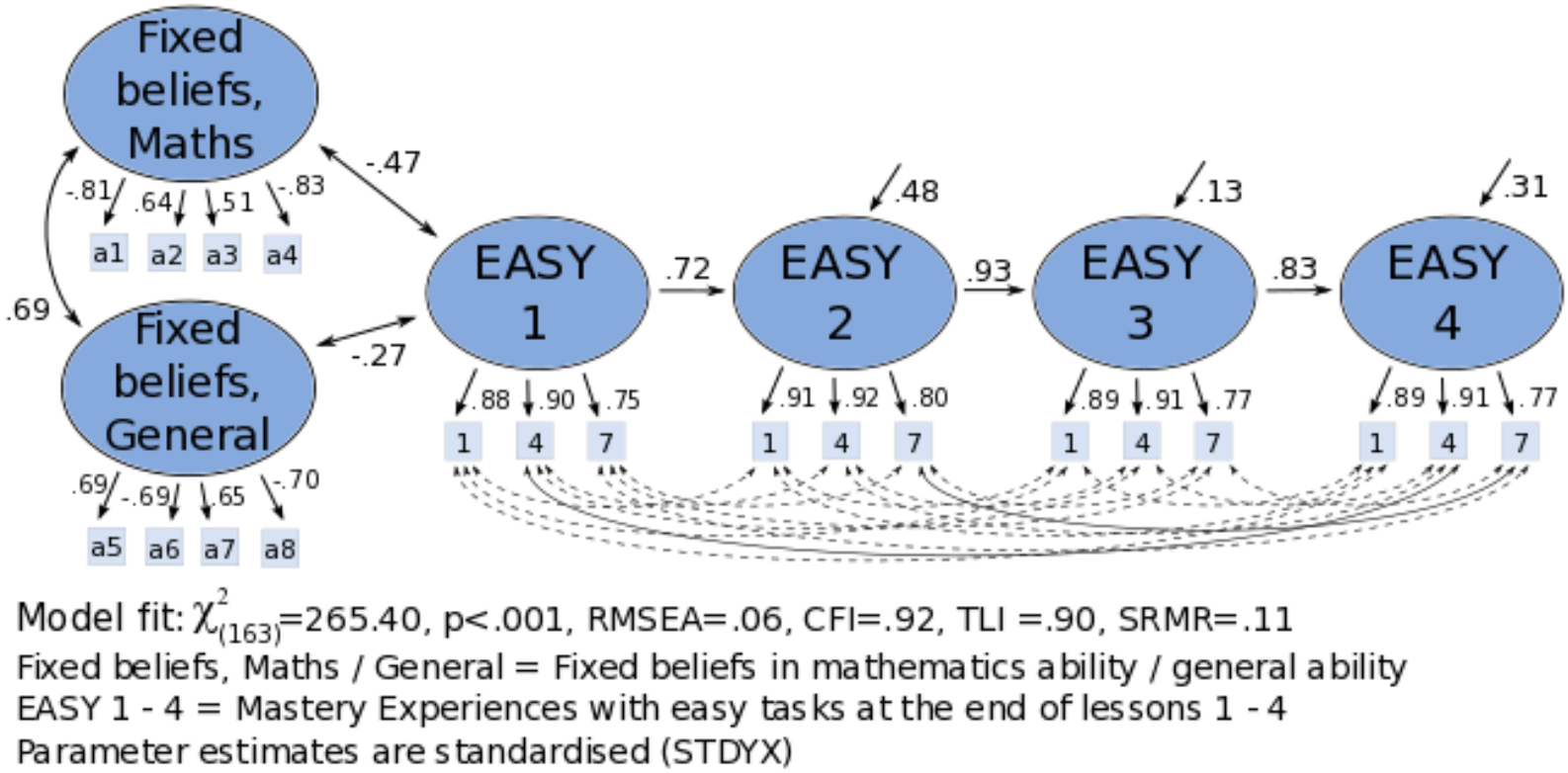
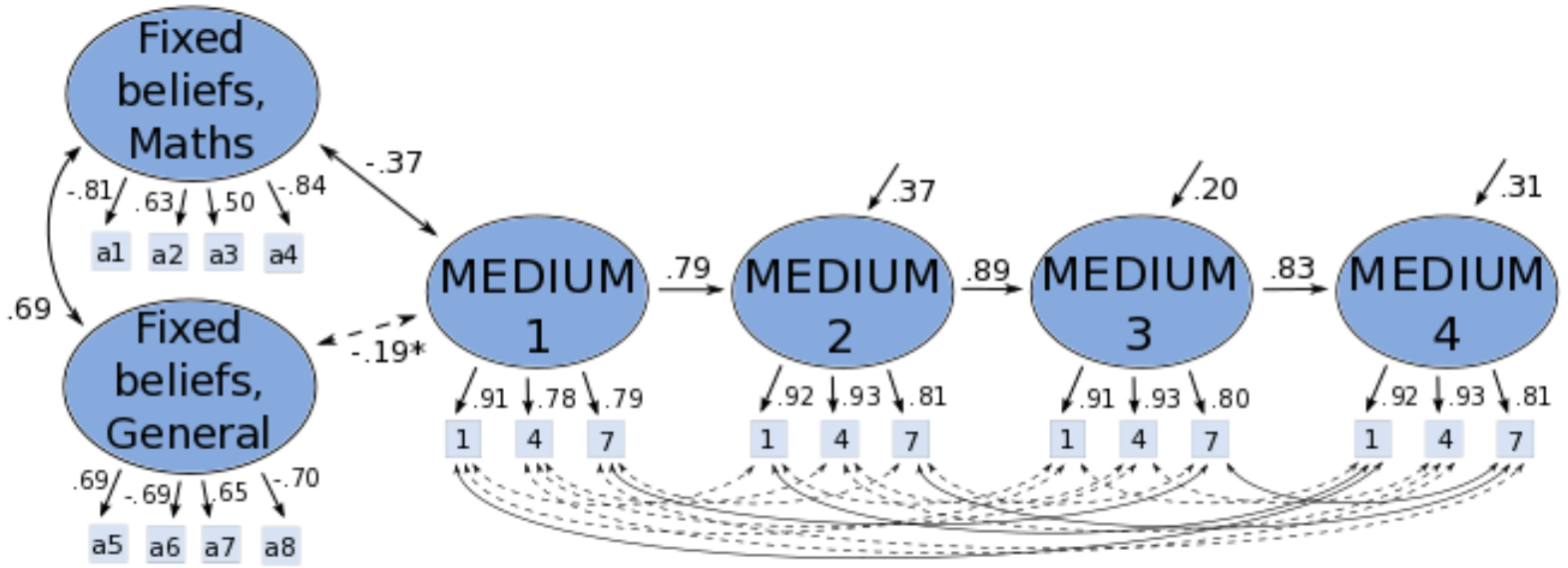


Figure 69: Simplex model of Mastery Experiences, easy tasks, with Ability Conceptions



*p = .105

Model fit: $\chi^2_{(162)}$=320.33, p<.001, RMSEA=.07, CFI=.89, TLI =.87, SRMR=.10

Fixed beliefs, Maths / General = Fixed beliefs in mathematics ability / general ability

EASY 1 - 4 = Mastery Experiences with medium difficulty tasks at the end of lessons 1 - 4

Parameter estimates are standardised (STDYX)

Figure 70: Simplex model of Mastery Experiences, medium difficulty tasks, with Ability Conceptions

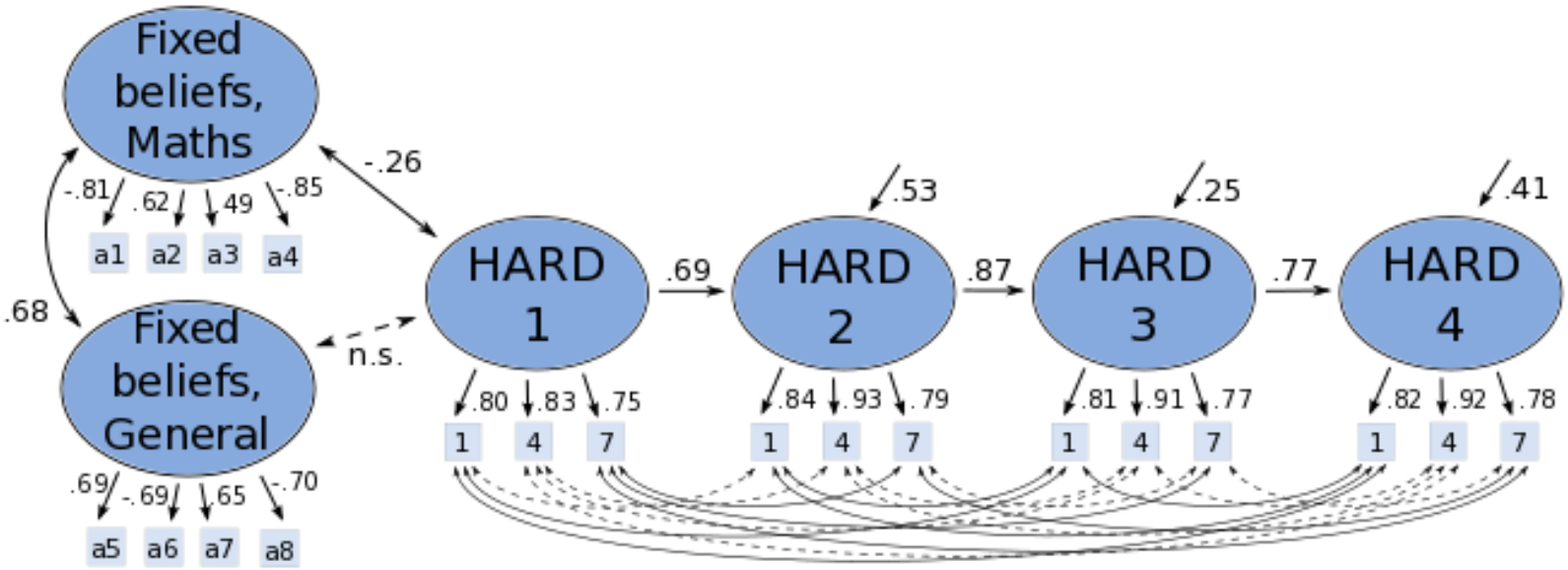


Figure 71: Simplex model of Mastery Experiences, hard tasks, with Ability Conceptions

Overall, the results from the mastery experiences models were comparable to those including self-efficacy. As can be seen in Figures 69 to 71, students' fixed beliefs in mathematics ability were significantly related to their mastery experiences on tasks of all levels, while the relationship between students' fixed beliefs in general ability and mastery experiences was significant in relation to easy tasks, marginally significant in terms of medium difficulty tasks, and non-significant for hard tasks. Similar to the self-efficacy construct, the strength of the relationship between fixed beliefs in ability and mastery experiences seemed to progressively weaken with increasing task difficulty. Again, significant relationships with constructs adapted from an established measure for ability conceptions supported the validity of my proposed measure for mastery experiences.

5.4.3.5 The relationship between ability conceptions and task self-efficacy

For the sake of consistency, I tested also the relationship between ability conceptions and Task self-efficacy in a path model, based on the autoregressive model developed in section 5.4.2.2. As can be seen in Table 99, including ability conceptions as covariates in the autoregressive model for students' Task self-efficacy did not result in very good fit – in fact, most of the fit indices failed to reach an "acceptable" level. As I do not have a strong theoretical reason to propose a specific alternative model, and the modification indices were inconclusive, I chose not to re-specify this model. The parameter estimates from this model indicated that neither type of fixed beliefs was significantly related to students' Task self-efficacy at the end of lesson 1.

Table 99: Results Task Self-Efficacy model with Ability Conceptions as covariates

	Fit indices					
Model	χ^2	*p*	RMSEA	CFI	TLI	SRMR
Task Self-Efficacy, simplex model*, equality constraints on lesson 2-4 paths. Lesson 1 Task self-efficacy ON ability conceptions	$\chi^2_{(52)}$= 132.19 χ^2/df= 2.54	p<.001	0.09	0.84	0.80	0.11

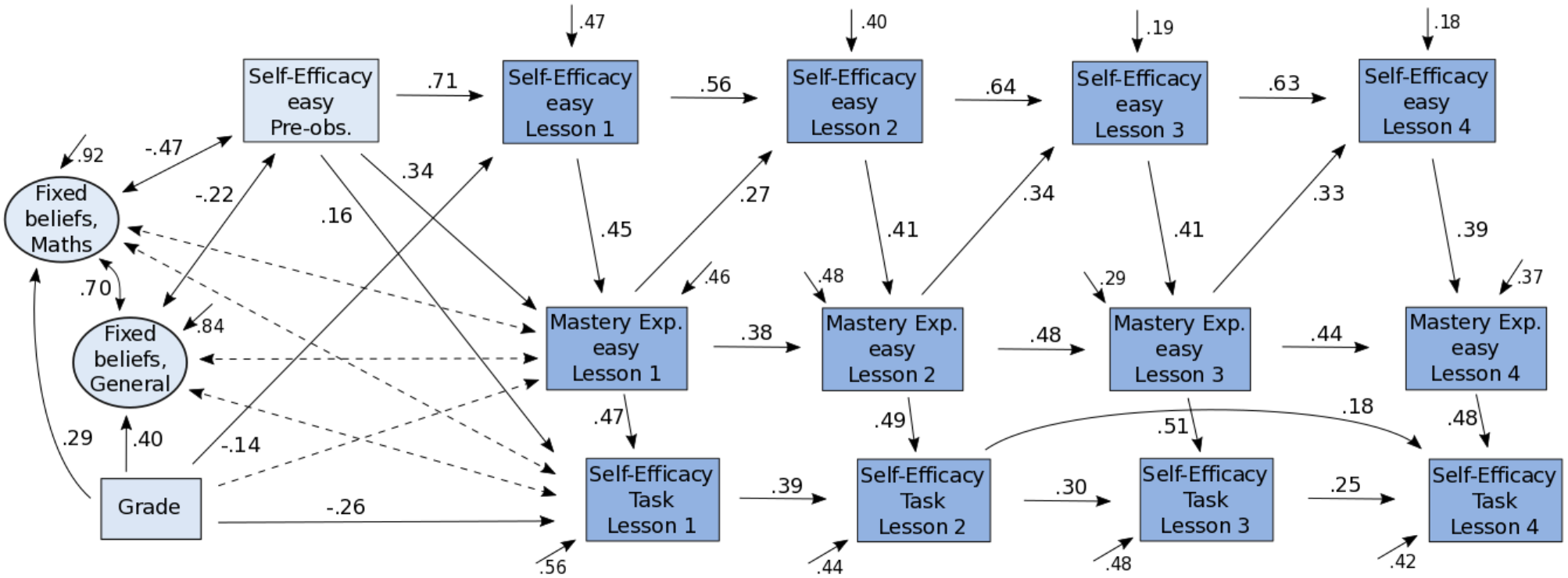
*including a path from lesson 2 to lesson 4

5.4.3.6 The relationship between ability conceptions, self-efficacy, and mastery experiences

On the basis of the cross-lagged path model developed in research question 2, section 5.4.2.3, I specified for each of the three level constructs a model where students' fixed beliefs in mathematics ability and general ability were correlated with their self-efficacy, mastery experiences, and task self-efficacy at lesson 1 (Models 1, 2, and 3 in Table 100). The parameter estimates of these models were not meaningful, as they indicated there were no significant relationships between ability conceptions and either of the measures included, contrary to the findings presented in sections 5.4.3.3. and 5.4.3.4. I consequently re-specified these models, where students' self-efficacy at pre-observations, rather than at lesson 1, were correlated with their ability conceptions (Table 100, Models 1b, 2b, and 3b, respectively, for each of the three level constructs). Models 1b, 2b, and 3b resulted in somewhat improved fit indices and meaningful parameter estimates, which are displayed in Figures 72 to 74.

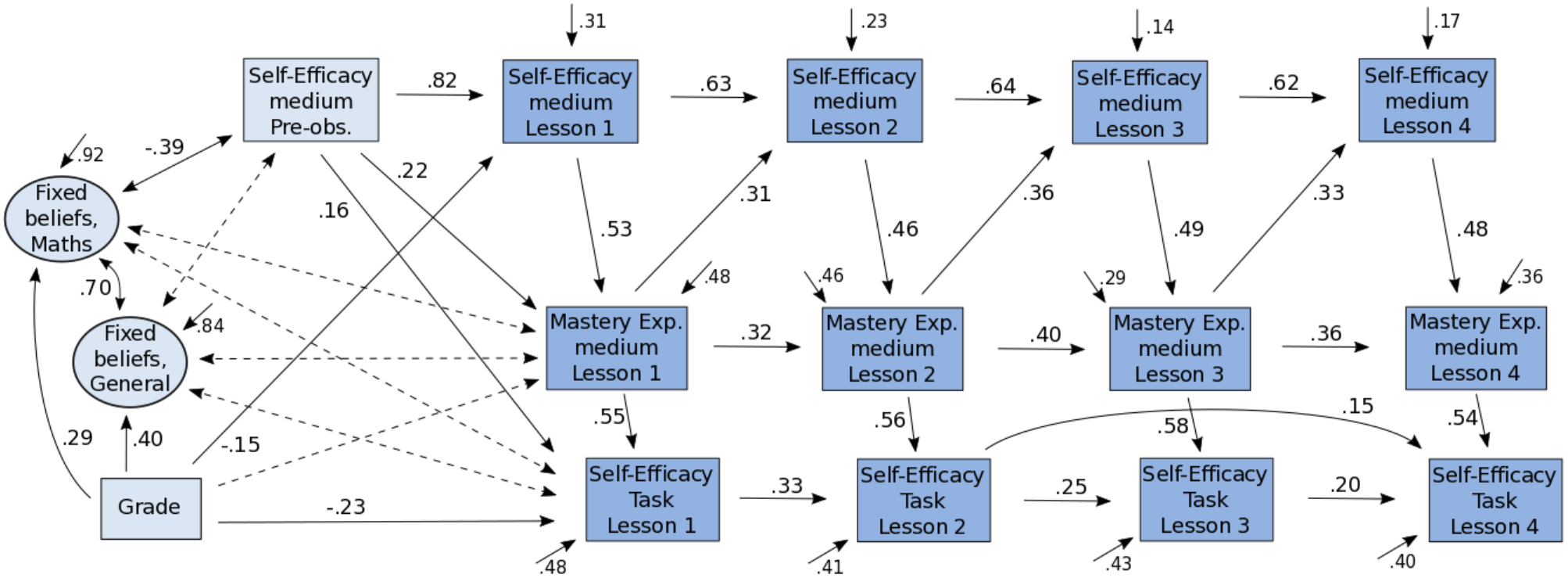
Table 100: Fit indices of path models of Self-Efficacy, Mastery Experiences, Task Self-Efficacy, with Ability Conceptions as covariates

		Fit indices					
	Model	χ^2	*p*	RMSEA	CFI	TLI	SRMR
1	Full path model including easy level constructs, regressed on grade and pre-obs., correlated with fixed beliefs	$\chi^2_{(193)}$= 376.37 χ^2/df= 1.95	p<.001	0.08	0.89	0.87	0.15
1b	Full path model including easy level constructs, regressed on grade and pre-obs. Pre-obs. correlated with fixed beliefs	$\chi^2_{(194)}$= 372.33 χ^2/df= 1.92	p<.001	0.07	0.90	0.88	0.11
2	Full path model including medium level constructs, regressed on grade, pre-obs., and fixed beliefs	$\chi^2_{(193)}$= 368.76 χ^2/df= 1.91	p<.001	0.07	0.91	0.89	0.13
2b	Full path model including medium level constructs, regressed on grade and pre-obs. Pre-obs. correlated with fixed beliefs	$\chi^2_{(194)}$= 353.68 χ^2/df= 1.82	p<.001	0.07	0.92	0.91	0.09
3	Full path model including hard level constructs, regressed on grade, pre-obs., and fixed beliefs	$\chi^2_{(193)}$= 314.49 χ^2/df= 1.63	p<.001	0.06	0.93	0.92	0.10
3b	Full path model including hard level constructs, regressed on grade and pre-obs. Pre-obs. correlated with fixed beliefs	$\chi^2_{(194)}$= 311.57 χ^2/df= 1.61	p<.001	0.06	0.94	0.93	0.10



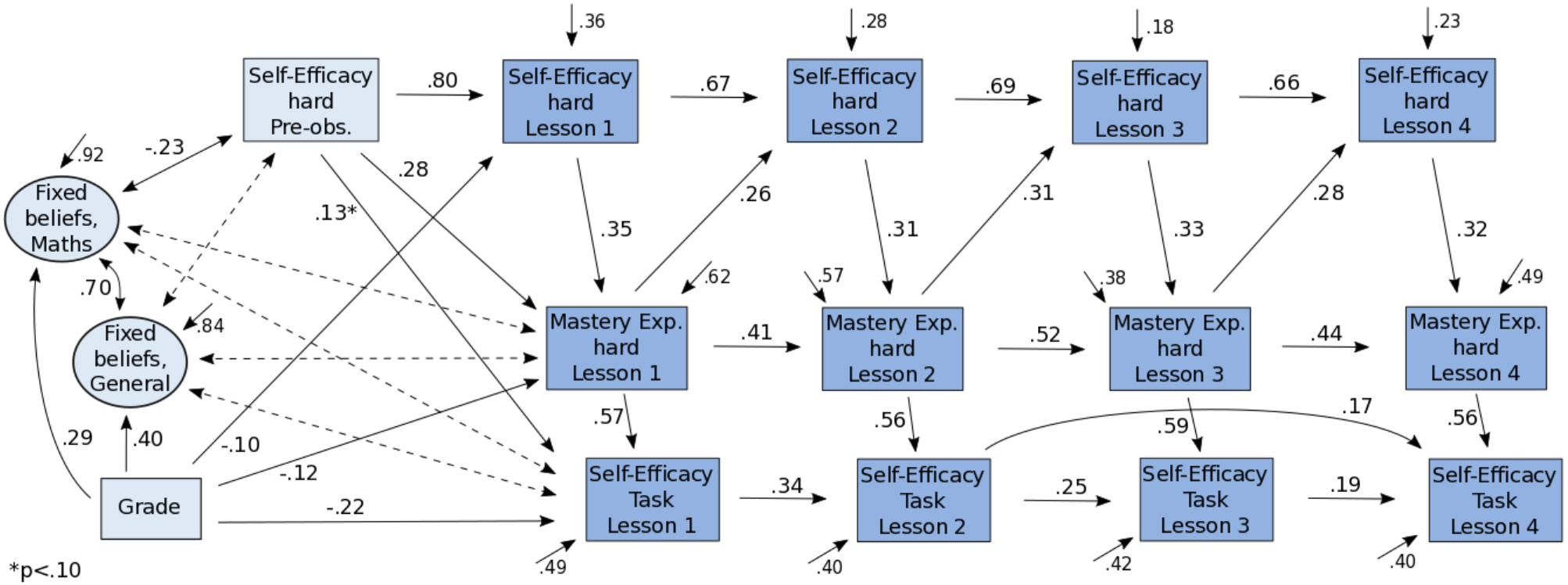
Model fit: $\chi^2_{(194)}$=372.33, p<.001, RMSEA=.07, CFI=.90, TLI =.88, SRMR=.11
Pre-obs. = Pre-observations, Mastery Exp. = Mastery Experiences, Fixed beliefs, Maths / General = Fixed beliefs in mathematics ability / general ability
For the sake of clarity, the fixed beliefs measurement models are not included.
Parameter estimates are standardised (STDYX). Dotted paths are estimated, but non-significant.

Figure 72: Cross-lagged path model of Self-efficacy and Mastery Experiences, easy tasks, with Ability Conceptions



Model fit: $\chi^2_{(194)}$=353.68, p<.001, RMSEA=.07, CFI=.92, TLI =.91, SRMR=.09
Pre-obs. = Pre-observations, Mastery Exp. = Mastery Experiences, Fixed beliefs, Maths / General = Fixed beliefs in mathematics ability / general ability
For the sake of clarity, the fixed beliefs measurement models are not included.
Parameter estimates are standardised (STDYX). Dotted paths are estimated, but non-significant.

Figure 73: Cross-lagged path model of Self-efficacy and Mastery Experiences, medium difficulty tasks, with Ability Conceptions



*p<.10

Model fit: $\chi^2_{(194)}$=311.57, p<.001, RMSEA=.06, CFI=.94, TLI =.93, SRMR=.10
Pre-obs. = Pre-observations, Mastery Exp. = Mastery Experiences, Fixed beliefs, Maths / General = Fixed beliefs in mathematics ability / general ability
For the sake of clarity, the fixed beliefs measurement models are not included.
Parameter estimates are standardised (STDYX). Dotted paths are estimated, but non-significant.

Figure 74: Cross-lagged path model of Self-efficacy and Mastery Experiences, hard tasks, with Ability Conceptions

As can be seen in Figures 72 to 74, for all levels of task difficulty, students' fixed beliefs in mathematics ability were (negatively) related to their mathematics self-efficacy at pre-observations. Students' fixed beliefs in general ability were (negatively) related to their mathematics self-efficacy for easy tasks only. Furthermore, grade level predicted students' fixed beliefs in ability, where grade 10 students were more likely to hold fixed beliefs in both mathematics ability and general ability, as compared with grade 6 students. The significant relationships between ability conceptions and mastery experiences I found when these constructs were modelled on their own (see section 5.4.3.4) were no longer evident when self-efficacy was also included in the models. In the present models, the relationship between fixed beliefs in ability and students' mastery experiences can be traced via the relationships with students' self-efficacy at the pre-observation time point, for each of the three level constructs.

Consistent with the findings in sections 5.4.3.3 and 5.4.3.4, the relationship between students' fixed beliefs in ability and their self-efficacy was strongest in relation to easy level tasks (-.47 and -.22 for fixed beliefs in mathematics ability and general ability, respectively), and progressively weaker for increasing task difficulty (-.39 for medium difficulty tasks and -.23 for hard tasks in the case of fixed beliefs in mathematics ability; the relationship with fixed beliefs in general ability was non-significant). These differences were not tested for statistical significance, as the three level constructs were specified in separate models. Furthermore, for all levels of task difficulty, grade level negatively predicted students' self-efficacy both through direct effects on self-efficacy at the beginning of lesson 1 and Task self-efficacy at the end of lesson 1, as well as indirectly through students' fixed beliefs in ability.

5.4.3.7 Summary of results for RQ3

In order to validate my proposed measures of self-efficacy and mastery experiences, I investigated their relationship with students' conceptions of ability in separate models. In the case of both of my multidimensional measures (the SEGD and the MEGD) significant relationships were confirmed in regard to students' fixed beliefs in mathematics ability and general ability, which served to validate my proposed constructs. In an absolute sense, the relationships with self-efficacy and mastery experiences were stronger in the case of fixed beliefs in mathematics ability as compared with fixed beliefs in general ability. Furthermore, they were stronger in relation to easy tasks, and progressively declined in strength with increasing task difficulty (again, in an absolute sense). These tendencies were not tested for statistical significance. Finally, I found no significant relationships between students' ability conceptions and their Task self-efficacy at the end of lesson 1.

In order to further investigate the relationship between students' ability conceptions and their self-efficacy and mastery experiences, I specified three cross-lagged path models (for easy, medium difficulty, and hard tasks) on the basis of the path models developed in research question 2, section 5.4.2.3. These models demonstrated that grade level predicted students' ability conceptions, where grade 10 students were more likely to hold fixed beliefs about ability as compared with grade 6 students. For all three levels of difficulty there was a significant relationship between students' fixed beliefs in mathematics ability and their self-efficacy, while no significant relationships with students' mastery experiences were found. In the case of students' fixed beliefs in general ability there was a significant relationship with their self-efficacy for easy tasks only. In an absolute sense, the relationship between students' fixed beliefs in ability and their self-efficacy decreased with increasing task difficulty.

5.5 Discussion

In the present study I set out to investigate the reciprocal relationship between students' self-efficacy and mastery experiences across a sequence of lessons in mathematics. I also investigated the relationship between students' self-efficacy and mastery experiences and their conceptions of ability, in order to validate my proposed latent level constructs. Consistent with study 3 I used a micro-analytic framework to investigate autoregressive and cross-lagged paths across a sequence of lessons in mathematics, as students were introduced to new mathematical topics. In the following sections I will discuss the key theoretical and practical implications from my findings, organised by research question.

5.5.1 A multidimensional measure of mastery experiences from lessons in mathematics

In research question 1 confirmatory factor analyses provided empirical support, consistent with the expectation, for the structural validity of my proposed mastery experiences measure (the MEGD), which was developed to match my self-efficacy measure (the SEGD), investigated in study 3. Consistent with my findings in study 3, my analyses demonstrated that a factor structure including three latent constructs for easy, medium difficulty, and hard tasks, and correlated uniquenesses representing facets of mathematics, fit my data best. Furthermore, tests of invariance supported strong or partially strong invariance of my measure across time, supporting that it was reasonable to use this to investigate changes in students' mastery experiences across a sequence of lessons.

To my knowledge, no previous research has proposed matched instruments for students' self-efficacy and their mastery experiences. In contrast, self-efficacy items are commonly matched with actual performance tasks, an approach introduced by Dowling (1978) with the Mathematics Confidence Scale (see section 2.2.1). The distinction between "actual" performance and mastery experiences is an important one, given the theoretical proposition

that performances do not influence students' self-efficacy directly, but are cognitively appraised. I argue that in order to increase our understanding of the theoretically proposed *process* of self-efficacy change, measures of mastery experiences (e.g. cognitively appraised performance experiences) are more appropriate to use, whereas measures of actual performance are better suited when the aim is to investigate performance as predicted by self-efficacy.

Similar to my argument in section 4.5.2, including levels of perceived difficulty in the measure of students' mastery experiences emphasises a theoretical position of task difficulty as integral to, and not separate from, students' judgments of success. For teachers, this implies that students differentiate between their experiences with different level of difficulty tasks. Students may consider task difficulty in terms of how challenging the task is to themselves (e.g. how much effort is needed), or in relative terms (e.g. how many students succeed on the task). Schweinle et al. (2013) argued that tasks framed according to "relative" difficulty might threaten students' self-efficacy. Accordingly, teachers might want to emphasise task difficulty in terms of personal challenge, and support students' accurate appraisal of what they can accomplish with strategic effort, in order to support their self-efficacy. Focusing on the skills, strategies, and effort required to complete tasks successfully is likely to lead students' attention beyond *Can I do this?* toward *How can I do this?* (Schweinle et al., 2013). A further implication stemming from my proposed matched instrument is that researchers can investigate empirically the reciprocal relationship between self-efficacy and mastery experiences for tasks of different levels of difficulty, as I did in research question 2 of study 4.

5.5.2 Reciprocal relationships between self-efficacy and mastery experiences

In research question 2, I included the best-fitting autoregressive models for my matched self-efficacy and mastery experiences measures, as well as students' Task self-efficacy, in a cross-

lagged path model to investigate reciprocal effects across a sequence of lessons. Similar to my expectation, all autoregressive and cross-lagged paths were significant, for easy, medium difficulty, and hard level tasks. Similar to the rank-order of students' self-efficacy investigated in study 3, section 4.4.3, the stability of the rank-order of students' mastery experiences did not change significantly across time. Furthermore, the cross-lagged paths between self-efficacy and mastery experiences did not differ significantly over time, which enabled statistical comparisons of the relative strength of autoregressive (e.g. construct stability) as well as cross-lagged paths (e.g. effect of one construct on another). My analysis gave rise to the following key findings:

1) students' self-efficacy was significantly more stable than their mastery experiences, in the case of medium difficulty and hard tasks;
2) stability and cross-lagged paths of lesson self-efficacy as well as mastery experiences did not change over time;
3) a reciprocal effects model was supported, where stronger self-efficacy predicted larger increases in mastery experiences, and more successful mastery experiences predicted larger gains in self-efficacy;
4) the dominant cross-lagged effect was from lesson self-efficacy to mastery experiences, where this difference was significant in the case of easy and medium difficulty tasks; and
5) there were differences in magnitudes of the effects according to task difficulty.

I discuss each of these findings below.

The first key finding, concerning the higher stability of students' self-efficacy as compared with their mastery experiences, was unsurprising, given that mastery experiences are formed on the basis of actual performance experiences. Social cognitive theory posits that students'

self-efficacy is influenced by a total of four sources, all of which are weighted and cognitively appraised (Usher & Pajares, 2008). It is likely that the kinds of explanations students make use of when they interpret effort, ability, and performances (e.g. their cognitive appraisal "filters") have built up over time and have stabilised, as per the research by Nicholls (1978). By way of contrast, mastery experiences reflect appraisals of students' performance experiences "only", where their performance experiences are influenced by several factors uncontrollable to students (e.g. topic, task difficulty, and task support). It is possible that *actual* performance is less stable than *interpretations* of performance, and that students' self-efficacy is more stable because it is one further step removed from students' actual performance, as compared with their mastery experiences. In short I argue that cognitive appraisal is associated with stability, due to a tendency of confirmation bias, e.g. interpreting information in light of previous experiences and self-constructs. This is in line with my argument in study 3, section 4.5.3, where I referred to confirmation bias in self-efficacy as discussed by Butz and Usher (2015). I discussed how teachers' support of students' interpretations of their experiences might be as important as facilitating positive performance experiences in the first place. The findings from study 4 provide further empirical support for this interpretation, through demonstrating the relatively higher stability of students' self-efficacy as compared with their mastery experiences. Future studies should investigate the hypothesis that rank-order stability is highest for students' self-efficacy, followed by their mastery experiences, followed by their actual performances.

Still in relation to the first key finding, the rank-order of students' self-efficacy remained highly stable, even as the rank-order of the same students' mastery experiences changed to a higher degree (for medium difficulty and hard tasks). Considering mastery experiences are viewed as the most powerful source of subsequent self-efficacy (see e.g. Byars-Winston et al.,

2017), this has implications for practice, as it highlights the central role of students' cognitive appraisals. If changes to the rank-order of students' mastery experiences are not reflected by comparable changes to the rank-order of their self-efficacy, it becomes important for teachers to ascertain how students typically appraise their experiences in mathematics. Teachers might need to work with a long-term goal to support adaptive interpretations, if students' cognitive appraisals cause them to view their performances in a negative light, or if students tend to focus on their performance relative to others, rather than in terms of their own task accomplishments. This might be especially pertinent in the case of students with low confidence in their mathematics abilities, as low self-efficacy is associated with maladaptive learning behaviours, while somewhat optimistic self-efficacy appears to be beneficial to students, particularly in western contexts (Klassen, 2004).

Previous studies have pointed to the role of the teacher in regard to students' beliefs about knowledge and ability (e.g. Cooper, 1979; Muis & Duffy, 2013; Snyder, Malin, Dent, & Linnenbrink-Garcia, 2014), which are likely related to students' cognitive appraisals. Muis & Duffy (2013) demonstrated that university students' academic self-efficacy changed across an intervention designed to influence the "instructional climate", where the professor used modelling, dialogue, and questions to support constructivist learning beliefs. Similar approaches might be effective in regard to changing the classroom climate of younger students, also. Considering the possibility of confirmation bias, my findings also highlight the importance of unambiguous and specific performance feedback, to lower the likelihood of student misinterpretation of messages conveyed by the teacher.

Considering the second key finding: The stability as well as cross-lagged paths between lesson self-efficacy and mastery experiences all remained stable across time. I will focus on

the cross-lagged paths, as I have discussed already the stability paths. According to theory, students' self-efficacy and performance should tend toward an equilibrium, across a certain amount of cycles (Lindsley et al., 1995; Shea & Howell, 2000; Williams & Williams, 2010). Shea & Howell (2000) found that the effect of self-efficacy on *performance* decreased in magnitude across a couple of cycles. Optimistic self-efficacy will not lead to continued increases in performance, which is limited also by current knowledge and ability (see section 5.2). In contrast to this, in my study the reciprocal effects in both directions could be constrained as equal without significant detriment to fit. This implies there are qualitative differences between the self-efficacy / mastery experiences relationship, as compared with the self-efficacy / performance relationship. While the influence from performance dwindles, the effect from mastery experiences on students' self-efficacy in my study was sustained – even as the growth of students' mean-level self-efficacy flattened (as per study 3, section 4.4.4.2). The most obvious explanation lays in the difference between the performance and the mastery experience constructs: student appraisals. If teachers direct attention to students' learning progress, this might support gains in their self-efficacy.

Considering the third key finding from research question 2; a reciprocal effects model was supported, where increases in self-efficacy predicted stronger mastery experiences, while stronger mastery experiences also predicted gains in self-efficacy. This finding provided important empirical support for reciprocal effects between self-efficacy and mastery experiences with school age children. My finding demonstrates that (contrary to findings from previous research) stronger self-efficacy predicts increases in mastery experiences, indicating that students' mastery experiences can be affected by way of raising their self-efficacy. Similarly, stronger mastery experiences predict gain in students' self-efficacy, demonstrating

that as students learn a new topic and make mastery appraisals, their self-efficacy tends to also increase.

I will discuss key finding three together with key finding four: Contrary to my expectation, the effect from students' beginning-of-lesson self-efficacy to their end-of-lesson mastery experiences was stronger than the effect from their mastery experiences to their self-efficacy at the beginning of the subsequent lesson (in the case of easy and medium difficulty tasks). Previous research (see Talsma et al., 2018) has provided empirical evidence for a dominant effect in the opposite direction, from performance to self-efficacy, similar to my own findings in study 2, and non-significant effects from self-efficacy to performance in the case of school-age children.

My results from study 4 are interesting, and I propose three possible reasons for the differences between my current findings and the previous studies. First and foremost, my present findings pertain to students' mastery experiences, while most previous research (e.g. Talsma et al., 2018) and my own study 2 included measures of actual performance. As I pointed out above, there are important distinctions between students' performances and their mastery experiences. Second, my matched measures of self-efficacy and mastery experiences ensured high correspondence, while it is possible some previous studies have failed to find significant effects between self-efficacy and performance due to a lack of congruence between the measures (Klassen & Usher, 2010; Pajares & Miller, 1995). Third and final, Talsma et al. (2018) reported that longer lags between measurement occasions were associated with weaker cross-lagged effects. In my study the (stronger) cross-lagged effect from self-efficacy to mastery experiences pertained to a shorter time lag (generally 45-60 minutes) than the (weaker) effect in the opposite direction, which generally spanned across

one or two days. It will be important in future studies to unpack the relative contribution of construct stability versus measurement lag on different reciprocal effects between self-efficacy and mastery experiences.

The stronger effect of students' self-efficacy on their mastery experiences (fourth key finding), rather than vice versa, might reflect the relative stabilities of the constructs, as discussed above, where the effect of self-efficacy on mastery experiences is stronger because self-efficacy is the more stable construct. Even so, in a reciprocal change framework, I would argue that in order to change self-efficacy, changing the least stable construct in the process might offer the best chances of success. Accordingly, the path from mastery experiences to self-efficacy might be a reasonable starting point for teachers. Importantly, given the surprisingly high stability of self-efficacy even as students engage with new tasks (as discussed in section 4.5.3, study 3), it is possibly not enough for teachers to "merely" support students' successful task performances, to affect changes in the rank-order of students' self-efficacy. If this were the case (e.g. if successful task performance led to a higher rank-order standing in terms of self-efficacy), I would have expected somewhat lower stability of the rank-order in study 3, reflecting different rates of learning a new task. It is possible that the cyclical nature of mathematics, where new topics are still reliant on knowledge accumulated from previous topics to a high degree, causes high stability even when the topic is meant to be new to the students. Also, it is possible that high stability of cognitive appraisals cause higher stability to self-efficacy than would otherwise be the case (if actual performance, rather than cognitively appraised performance, directly changed self-efficacy).

Educators might want to consider that students cognitively appraise their performances in a way that their interpreted performances (e.g. mastery experiences) may be more stable than

their actual performance. As such, students might end up in a maladaptive cycle of self-efficacy / mastery experiences, where even successful "actual" performances do not affect positive changes in their self-efficacy. This implies that an integrated approach, where teachers focus on student cognitive appraisal (i.e. student sense-making of performance experiences) as much as they work to facilitate positive performance experiences, might be necessary to support changes in students' self-efficacy. The effect of this integrated approach would hopefully support students' mastery experiences, whilst also strengthening the effect from students' mastery experiences on their self-efficacy.

My fifth and final key finding from research question 2 was in relation to the role of task difficulty in regard to differences in construct stability and cross-lagged effects. My findings indicated that the rank-order of students' self-efficacy was progressively more stable with increasing task difficulty. In contrast, the stability of students' mastery experiences was lowest, and cross-lagged effects between the constructs were highest, in relation to medium difficulty tasks. These differences were in absolute terms only (as the constructs were investigated in separate models according to levels of difficulty), but are nonetheless interesting and worth discussing. If the stability of self-efficacy increases with task difficulty this means students are consistently confident (or consistently lack confidence) about their performance capabilities on hard tasks, while the rank-order of their confidence varies more with easier tasks. This might reflect a student perception that performance on hard tasks is determined by more stable factors such as ability, while performance on easier tasks is less stable, according to, for instance, situational effort. In contrast, the stability of mastery experiences was about similar for easy and hard tasks, and the least stable in relation to medium difficulty tasks. Mastery experiences are students' interpretations of their performances, and this pattern of stability might reflect that performance on easy and hard

tasks is consistently easy / hard, while performance on medium difficulty tasks is more variable.

The cross-lagged effects of lesson self-efficacy and mastery experiences were strongest in relation to medium difficulty tasks, for which students' mastery experiences were the least stable. I believe that the variability of performance on medium difficulty tasks seems to provide a useful context for changing students' self-efficacy. It is likely that perceived medium difficulty tasks are difficult enough that students' self-efficacy is raised as an effect of successful task performance, however not so difficult that students think effort has little bearing on the outcome. Previous research has pointed to the role of tasks that are challenging, but solvable through effort (e.g. Pajares, 2002; Stylianides & Stylianides, 2014; Yetkin Özdemir & Pape, 2013) as important for improving students' self-efficacy. Bandura & Wood (1989) found a significant relationship between perceptions of controllability and self-efficacy, where a high sense of *controllability* was associated with higher intercept and growth of self-efficacy. Different to my findings, Stajkovic and Luthans (1998) reported that the strongest effects between self-efficacy and performance were found in relation to easy tasks. These differential findings might be related to the differences between *objectively* measured task difficulty, as per the study by Stajkovic & Luthans (1998), versus *perceived* levels of difficulty, as investigated in my study. On the one hand, it is possible that high stability of self-efficacy and mastery experiences for hard level tasks is associated also with confirmation bias, where students with low self-perceptions interpret successful performance on hard tasks in terms of luck; in other words, the students do not "believe" successful performance is indicative of future performance potential. On the other hand, successful performance on easy tasks might be seen as unimportant, e.g., the task was easy, thus successful performance may convey little information about future performance capability. In

contrast, performance on "adequately" challenging tasks might provide students with both believable and important information about their potential to perform on future tasks.

While my findings indicate the rank-order of students' self-efficacy is most likely to change as a result of engaging with medium difficulty tasks, the benefits from changing self-efficacy for hard tasks may be greater. It is possible that teachers may affect positive self-efficacy change through spending time on one or several relatively challenging tasks (when first presenting a new topic), with adequate scaffolding to support both successful performance and appraisal of the performance, in order to demonstrate to students that their perseverance can pay off. At the same time this could be a "high-risk" activity, where students with low self-efficacy might choose to disengage with the task early on, losing out on a valuable opportunity to experience mastery. Stylianides and Stylianides (2014) emphasised how teacher support was crucial at times when students might otherwise give up, which further demonstrates the careful balance act teachers need to strike, in order to provide appropriate support for students when they engage with tasks of different levels of difficulty. Examples of teacher scaffolding are encouraging students not to give up, asserting that the tasks are solvable through effort, providing hints or tips to support successful mastery, and highlighting student progress relative to previous performances (Pajares, 2002; Stylianides & Stylianides, 2014; Yetkin Özdemir & Pape, 2013).

Teachers are limited in resources, and have to cater for students of differing abilities and self-efficacy in the classroom. They may want to employ a range of strategies for supporting students' self-efficacy, at different times. For instance, during "everyday" classroom work, teachers might focus on ensuring students work on appropriately challenging tasks (e.g. tasks of "medium difficulty"), provide timely and specific feedback to support students' cognitive

appraisals of their experiences on these tasks, model incremental beliefs in mathematics ability, and teach students how to monitor their own progress (e.g. improve the quality of the internal feedback mechanism to include relevant factual information, as per Lindsley et al., 1995).

In regard to changing the self-efficacy of students who are in a maladaptive cycle, more effort might be required than can be afforded by a single teacher on a daily basis. Again, starting with mastery experiences is likely more effective than trying to otherwise change self-efficacy directly. Nevertheless, it is possible that the "everyday" tasks students engage with during typical classroom lessons are not sufficient to cause changes if students have built up a stable, negative self-perception of their ability. Previous research has pointed to the role of tasks that are meaningful and memorable, as well as of appropriate cognitive demand (e.g. Stylianides & Stylianides, 2014) in order to change maladaptive student beliefs in mathematics. If such beliefs are ingrained it is likely that careful planning and task implementation is necessary in order to ensure not only successful student performance, but also adaptive cognitive appraisal of the performance, and finally support to recognise the meaning of the performance in regard to self-efficacy for future mathematical tasks.

Future research into the cross-lagged effects and stability of self-efficacy and mastery experiences for different levels of task difficulty would be fruitful, as this could increase our theoretical understanding of the role of levels of difficulty in the process of self-efficacy change, and provide further implications for classroom practice. For now, I conclude that working on medium difficulty tasks seems to be beneficial to student self-efficacy change. Promoting "adequately challenging" tasks during classroom work in mathematics is hardly

controversial, but understanding how and why experiences on such tasks might contribute to adaptive self-efficacy is important in order for teachers to appropriately support this process.

5.5.3 Significant relationships with ability conceptions

In research question 3 I used an adapted version of a previously established measure of ability conceptions to validate my proposed measures of self-efficacy and mastery experiences, and to investigate the relationship between students' ability conceptions and their self-efficacy and mastery experiences for different levels of task difficulty. The key findings from my analysis were the following:

1) significant relationships between students' fixed beliefs in ability and their self-efficacy and mastery experiences supported the validity of my proposed measures;
2) there were significant relationships between students' self-efficacy and their ability conceptions in a cross-lagged path model including also mastery experiences; and
3) I found differences in the relationships between ability conceptions and self-efficacy / mastery experiences according to task difficulty.

I will now discuss each of these key findings.

In my first key finding from research question 3, autoregressive path models demonstrated significant relationships between students' fixed beliefs in mathematics ability and general ability, and their self-efficacy and mastery experiences. These results were consistent with my expectations and provided support for the validity of my proposed measures, through demonstrating expected relationships with a previously established measure. I interpret these findings in line with theoretical tenets (see section 5.2), where ability conceptions might function as a form of cognitive appraisal (among other factors / perceptions), and thus interact with the relationship between self-efficacy and mastery experiences. Students with high fixed beliefs in ability are likely to interpret performance experiences as indicative of their own

(fixed) ability, while students with low fixed beliefs in ability are likely to see performance as indicative of less stable factors, such as for instance effort, support, or luck (Dweck & Leggett, 1988).

Due to these different types of cognitive appraisals, students with high versus low fixed beliefs in ability are likely to interpret unsuccessful performance experiences very differently. For students with high fixed beliefs, "failure" is indicative of low ability, presumably associated with a decrease in self-efficacy. In contrast, experiences of "failure" to students with low fixed beliefs is likely less damaging to their self-efficacy, as these experiences are not construed as a sign of low ability. In regard to mastery experiences, a significant relationship between mastery experiences and ability conceptions probably reflects that students with higher fixed beliefs in ability do not "believe" that increased effort can lead to enhanced performance, and as a consequence are less likely to invest effort when engaging with mathematical tasks. This again results in lower performances and mastery experiences.

My second key finding from research question 3 was related to the relationships between students' ability conceptions, self-efficacy, and mastery experiences. While students' ability conceptions were significantly related to both self-efficacy and mastery experiences when investigated in separate models, only the relationship with students' self-efficacy was significant when all constructs were included in the cross-lagged path models developed in research question 2. Given the significant reciprocal relationship between self-efficacy and mastery experiences, it is likely that the relationship between students' ability conceptions and mastery experiences was mediated through their self-efficacy. Future studies can investigate this proposition formally.

Furthermore, the strength of the relationship with students' self-efficacy differed for students' fixed beliefs in mathematics ability and general ability, where the strength of the association with fixed beliefs in mathematics ability was more than twice as strong as the association with fixed beliefs in general ability, in the case of easy tasks. Comparable tendencies were found in the models where self-efficacy and mastery experiences were investigated with ability conceptions separately (sections 5.4.3.3 and 5.4.3.4). The differences in strength were not surprising, given the subject specificity of my adapted measure of fixed beliefs in mathematics, similar to the subject specificity of my measures of mathematics self-efficacy and mastery experiences. Previous research has pointed to differences in correlations as an effect of the correspondence between self-efficacy and performance measures (e.g. Talsma et al., 2018), an effect that is likely to be relevant for other constructs. Different to their fixed beliefs in mathematics ability, students' fixed beliefs in general ability were not significantly related with their self-efficacy for medium difficulty or hard tasks, when these relationships were investigated in the cross-lagged path model.

Accordingly, my third key finding from research question 3 included differences in relation to levels of task difficulty. The relationship between students' fixed beliefs in mathematics ability and their self-efficacy was strongest (in absolute terms) in relation to easy tasks, and progressively weaker with increasing task difficulty. In other words, students' who saw ability in mathematics as fixed were likely to hold weaker self-efficacy: and this effect was strongest in relation to easy level tasks. Similarly, the strength of the relationship between students' fixed beliefs in general ability and their self-efficacy decreased with task difficulty: this was demonstrated in the path models where self-efficacy and ability conceptions were modelled on their own (section 5.4.3.3). The same trends were evident in regard to mastery experiences

and ability conceptions (section 5.4.3.4), where the strength of the relationship between students' mastery experiences and their ability conceptions decreased with task difficulty.

This resembles the findings by Stajkovic and Luthans (1998), who reported that task difficulty moderated the relationship between self-efficacy and performance, where the strongest effects were found in relation to easy tasks. The authors further argued that ability conceptions might play a role in this complex interplay. Similarly Bandura & Wood (1989) found an effect of (perceived) task difficulty on students' self-efficacy. Unlike these indications of task difficulty as a moderator of the self-efficacy / performance relationship, I am not aware of any previous research reporting that the strength of the relationship between ability conceptions and self-efficacy differs as a function of task difficulty. While I did not investigate these findings in my study for statistical significance, they are interesting and worth discussing, not the least considering the consistency of this trend across both self-efficacy and mastery experiences.

The stronger relationship between ability conceptions and self-efficacy / mastery experiences for easier tasks might be discussed in light of my findings in research question 2 in the present study, regarding the stability and change of self-efficacy and mastery experiences. I found that students' self-efficacy was the least stable in regard to easy level tasks, and progressively more stable with increasing (perceived) task difficulty. In other words students' self-efficacy for easy tasks is potentially the most amenable to change, and thus more readily influenced by self-efficacy changing information (e.g. mastery experiences, or cognitive appraisal in the form of ability conceptions).

An alternative explanation for the differential relationship between ability conceptions and self-efficacy / mastery experiences according to task difficulty concerns students' perceptions of the requirements of success on hard tasks, in relation to the extent to which ability can be influenced over a short period of time. As discussed in section 5.5.2, performance on hard tasks might to a larger degree be reliant on ability (in the form of either innate ability *or* present capability, depending on how ability is perceived by the individual) relative to effort, as compared with easier tasks. Even if ability is conceptualised as malleable, building up the level of ability required for successful performance on hard tasks might require practice over time. Accordingly, even if students have low fixed beliefs in ability (e.g. believing that ability is malleable) they may perceive that they do not *currently* possess the prerequisites for successful performance on hard tasks. This could explain why the relationship between ability conceptions and students' self-efficacy is weaker for harder tasks.

In regard to mastery experiences, I interpret the significant relationship between mastery experiences and ability conceptions in terms of students' effort, as discussed above. Similar to my interpretation above, if students perceive that successful performance on hard tasks requires a certain degree of ability, and they do not currently possess this ability, they may reduce their effort even if they do believe ability is malleable in the long run (e.g. have low fixed beliefs in ability). In contrast, low fixed beliefs in ability is likely to lead to increased effort in relation to easier tasks, resulting in stronger relationships between mastery experiences and ability conceptions for easier, as compared with harder, tasks.

If my above interpretation is correct, this would imply that working on tasks that are "too hard" might result in reduced self-efficacy and effort for all students, whether they tend to have fixed or malleable beliefs in ability. Bennet and Desforges (1988) demonstrated a

teacher tendency to underestimate high achievers and overestimate low achievers, resulting in ill-adapted tasks for both groups. In this view, the issue of "too hard" tasks is likely to be most prominent for low achieving students, or students who have low perceptions of their own ability in mathematics. If students give up and reduce their task effort, they lose out on valuable opportunities to learn and experience mastery on challenging tasks, which again is associated with reduced self-efficacy for these tasks. In order to avoid such maladaptive cycles, teachers might want to ascertain the appropriate level of task difficulty for each individual student in the classroom, and also provide scaffolding to, for example, highlight the role of effort and support successful performance.

If students perceive that performance on hard tasks is more reliant on ability than on easier tasks, then "evidence" of such ability may be relatively more important with increasing levels of perceived difficulty. This may be seen in light of my findings on research question 3, study 3 (see section 4.4.4.2) where latent growth curves demonstrated that students' self-efficacy for hard tasks grew more steeply after a single lesson experience, as compared with their self-efficacy for medium difficulty and easy tasks. This might be seen as an indication that students drew on their mastery experiences to a high degree when considering their self-efficacy for hard tasks.

In conclusion, students might attend differently to sources of self-efficacy for different levels of task difficulty. For instance, ability conceptions might play a relatively more important role in regard to self-efficacy for easier, as compared with harder tasks, while mastery experiences might be a relatively more important source in the case of harder, as compared with easier tasks. For teachers, this finding suggests that a range of strategies might be necessary in regard to supporting adaptive self-efficacy for students of differing abilities and self-

perceptions, working on different levels of difficulty tasks, across a variety of learning situations.

5.5.4 Further implications and limitations of my study

While some implications and limitations of my key findings from this study have been discussed according to each of my research questions, further implications and limitations of my study are worth noting. One is related to differential findings for students in grade 6 and grade 10. It is possible that cognitive appraisal is more stable for older students, as has been shown to be the case for other motivational constructs (e.g. Gottfried et al., 2001). I did not investigate differences in stability of self-efficacy and mastery experiences as a function of age in the current study, but this would be an important avenue for future research. Future studies should also investigate the stability of students' cognitive appraisals: while ability conceptions (as one factor likely influencing cognitive appraisal) were included in research question 3 of the present study they were measured at a single time point only, thus I could not investigate empirically my proposition that student cognitive appraisal is highly stable across time, and more stable than their self-efficacy.

Grade level was a significant predictor in both research question 2 and 3, where grade 10 students were more likely to hold weaker self-efficacy at lesson 1, as compared with the younger students. This is in accordance with previous findings (e.g. Bong, 2009). Grade level also predicted students' mastery experiences at lesson 1 (negatively), however only in regard to hard tasks. Finally, older students were more likely to hold fixed beliefs in mathematics ability and general ability, where of course fixed beliefs in mathematics ability was (negatively) related to subsequent self-efficacy. It is possible that grade level differences exist not only in regard to the stability and strength of students' self-efficacy and mastery experiences, but in regard also to the *relationship* between self-efficacy and mastery

experiences. Previous research has indicated older students are more adept at assessing their own capabilities as compared with younger students (Butz & Usher, 2015; Davis-Kean et al., 2008; Wigfield et al., 1997), which might interact with the mastery experiences – self-efficacy relationship. My findings in study 3 demonstrated significant differences in regard to mean intercept and growth of self-efficacy between students in grade 6 and grade 10, as well as differences in the relationship between intercept and growth. These differential findings were related to students' self-efficacy across a sequence of lessons engaging with a new topic, which implies there may well be a differential relationship between mastery experiences and self-efficacy for the two groups. Overall, I expect older students' self-efficacy and mastery experiences constructs may be more stable as compared with younger students (as per Wigfield et al., 1997), while the cross-lagged effects between lesson self-efficacy and mastery experiences are stronger in the case of younger students. Future studies can investigate these hypotheses formally.

Gender was not controlled in the current study, which could be another limitation. Previous studies have found that male students report stronger self-efficacy than females (e.g. Schulz, 2005; Pajares & Miller, 1994), and that girls and boys perceive and emphasise different sources of self-efficacy (e.g. Butz & Usher, 2015; Joët, Usher, & Bressoux, 2011). Gender stereotypical beliefs have been discussed as one possible explanation for such differences. It is possible that there are gender differences also in regard to the relationship between self-efficacy and mastery experiences, which would be a fruitful avenue for future research.

A further limitation of my study is that measures of previous performance were not included in the model. Given the reciprocal relationship between self-efficacy and performance (see e.g. section 3.5.1), this could have led to an overestimation of the effect from self-efficacy on

students' mastery experiences, particularly in relation to the first cross-lagged path in the model (lesson 1 self-efficacy to lesson 1 mastery experiences). It is likely that students' perceived mastery experiences at the end of lesson 1 would reflect in part their previous performance, thus alleviating this problem for the later cross-lagged paths.

Although the purposive sampling employed in phase 2 (see section 4.3.2.1) was aimed at including classes that were representative to the Norwegian grade 6 and grade 10 student population, there is a chance the participating classes differed to other Norwegian students. The modest sample size can thus be seen as a limitation of the study. Furthermore, although a high percentage of students in each class consented to participate, it was typical of most classes that one or several students did not follow the general mathematics lessons, due to for instance language barriers or a diagnosed learning disability. Thus, even to the extent the participating classes are representative to the Norwegian population, the findings do not extend to these students. It would be valuable in future studies to investigate the relationship between self-efficacy and mastery experiences with a larger sample, and to include investigations into whether this relationship differs for students who follow the ordinary mathematics lessons versus students that are routinely taken out of class.

Other implications stem from my measure of Task self-efficacy. My models demonstrated that, overall, students' lesson self-efficacy was more stable across time than their mastery experiences, which again were more stable than their Task self-efficacy. This is unsurprising, given that *tasks* are more specific, and presumably vary more, as compared with *lessons*. Furthermore, the Task self-efficacy measure consisted of a single item only, as compared with the multi-item lesson constructs. While including the single item measure provided interesting contrasts, it is hard to unpack whether differences between this construct and the other

constructs are methodological artefacts or reflect "real" substantive differences. Furthermore, the strongest cross-lagged effects in my models were from students' mastery experiences to their Task self-efficacy, both measured at the end of each lesson (in the case of easy and hard tasks; the strength of the effect was similar to the effect of lesson self-efficacy on mastery experiences in the case of medium difficulty tasks). This likely reflects an effect of measurement lag, as discussed in section 5.2.1, as the lag between the measures was in the realm of minutes. It may also reflect the fact the Task self-efficacy was related to confidence about performing on tasks similar to the main tasks of the last lesson, while the lesson self-efficacy was in relation to the next lesson, which students may (or may not) have expected to differ from the last lesson. In future studies it would be valuable to unpack the effects from facet-specificity of self-efficacy and measurement lag in relation to cross-lagged paths, from the effect of measurement approaches, in regard to both rank-order stability and cross-lagged effects.

Finally, when interpreting cross-lagged effects, it is crucial to consider the design, and, in particular the measurement lags, for each of the constructs. If either self-efficacy or mastery experiences are measured at a "suboptimal" lag, this might confound the results in terms of the effect of one appearing stronger than the other (and stronger than it actually is) (Little, 2013, p.48). There is scarce evidence from previous studies about the optimal lag for measuring reciprocal effects between self-efficacy and mastery experiences, thus I based my design decisions on theory, which asserts self-efficacy changes as an effect of influential events. Furthermore, I conceptualised mathematics lessons as a likely context for these influential events. A shorter interval (e.g. Bernacki et al., 2015) might be appropriate – in future studies a varied lag might illuminate the issue of what the optimal lag is for measuring changes in self-efficacy.

5.6 Conclusions

In the present study I set out to investigate the reciprocal relationship between students' self-efficacy and mastery experiences across a sequence of lesson learning a new topic in mathematics, as well as the relationship between self-efficacy, mastery experiences, and conceptions of ability. My findings provided support for a new multidimensional measure of mastery experiences, and for the validity of my measures of lesson self-efficacy (investigated in study 3) and mastery experiences. Furthermore, I found empirical support that students' self-efficacy was more stable than their mastery experiences, and that the effect of students' self-efficacy on their mastery experiences was stronger than the effect in the opposite direction, but mastery experiences did predict gains in self-efficacy in the subsequent lesson. I know of no previous studies that investigated the reciprocal relationship between students' self-efficacy and mastery experiences in mathematics across a sequence of classroom lessons. Furthermore, I found significant relationships between students' conceptions of ability and self-efficacy, where this relationship was weaker for harder, as compared with easier, levels of task difficulty. Accordingly, the findings provided empirical support to my proposition that levels of task difficulty are reasonable to consider and can contribute meaningfully to substantive investigations, through the indications of substantive differences in regard to stability as well as cross-lagged effects, according to levels of perceived task difficulty. Building on the findings from the first three studies of this thesis, my final study has provided important new knowledge regarding the development of students' self-efficacy in the classroom, with both theoretical and practical implications.

6 Overall discussion and conclusion

In the current thesis I set out to address two broad questions about students' self-efficacy and performance experiences with mathematics, through investigating Norwegian grade 5 through 10 students' self-efficacy and performance experiences in different settings and across differing time frames. In my introduction chapter I introduced the main themes of my thesis, which I will revisit in this final discussion chapter, and address in view of key findings from across my studies.

I started my introduction chapter posting two questions: What is the relationship between the kinds of experiences students have with mathematics, and their confidence that they can perform well on mathematical tasks? and: How does students' confidence for future learning and performance in mathematics develop during classroom lessons? I proposed a theoretical model of self-efficacy and students' experiences with mathematics (section 1.4), identified gaps in the literature in need of further investigation, and presented how I would contribute to knowledge by addressing these research gaps (section 1.5). Finally, I framed 11 research questions, which I addressed in four interrelated studies, presented in separate chapters of the thesis document. The four studies are related through the overarching questions, and the later studies built on the investigations in the previous ones. In the chapter for each study, I discussed the implications and limitations stemming from that investigation, thus in the current chapter I will not repeat the discussions for the individual research questions or studies, but rather focus on cross-study findings related to the main themes of the thesis.

In my thesis I set out to investigate the overarching theme of students' self-efficacy and experiences with mathematics (see section 1.3), with an aim to increase our knowledge about the relationship between students' self-efficacy and their performance experiences, and how students' mathematics self-efficacy changes over a sequence of classroom lessons when students learn new topics. I identified three important themes in relation to this overarching theme, which were a) the conceptualisation of self-efficacy, b) the relationship between self-efficacy and performance, and c) the development and change of self-efficacy. In the current chapter I will draw together and discuss contributions to theory from my thesis, with reference to these main themes. I will focus on findings across my four studies, which have contributed to knowledge through complementary foci: for instance the classroom learning versus test performance context; cross-sectional, longitudinal, or micro-longitudinal designs; and different methodological approaches. In order to facilitate this cross-study discussion, I have summarised the research questions and main findings from each of the studies in a table (Table 101). Each of the findings have been numbered in reference to the relevant study as well as the research question (e.g. findings from study 1, RQ 1, are assigned the numbering 1.1) as well as a letter (a, b, c etc.) to enable differentiation between key findings related to the same research question. In the following section I will present cross-study findings and implications, and discuss how they relate to the three important themes of my thesis (themes a, b, and c).

Table 101: Overview of research questions and findings from my thesis

Research questions for each study:		Findings from each study:		
	Study 1: Level, strength, and facet-specific self-efficacy, and students' performance on national tests in mathematics			
RQ 1:	*What is the structural validity of the proposed self-efficacy constructs, which include facets and levels of self-efficacy?*	1.1	a)	It was reasonable to measure students' mathematics self-efficacy with a multidimensional instrument including latent constructs representing levels of difficulty and test taking facets
RQ 2:	*How is facet-specific and level-specific self-efficacy associated with performance on national tests in mathematics?*	1.2	a)	There was a significant relationship between level of students' self-efficacy and their overall (mean) scores on national tests in mathematics. Stronger self-efficacy was related to higher national test scores for both grade 5 and 8/9.
			b)	Older students' self-efficacy was more strongly related to their performance scores than younger students' self-efficacy *
			c)	The relationship between self-efficacy and national test scores was strongest for self-efficacy related to medium levels of perceived difficulty
RQ 3:	*What is the relationship between students' self-efficacy for perceived easy, medium, and hard tasks, and the students' scores on easy, medium, and hard level items on national tests in mathematics?*	1.3	a)	Students' self-efficacy for each of the three levels of perceived task difficulty predicted their subsequent performance on corresponding task items
			b)	Students in grade 9 were most accurate at judging their capabilities, followed by the younger students
			c)	Students in grade 5 and grade 9 were most accurate at predicting their performances on easy and hard level tasks, while grade 8 students were most accurate in relation to medium level
			d)	Most students in my sample were overconfident, in relation to all task levels
	Study 2: Self-efficacy and mathematics performance: reciprocal relationships			
RQ 1:	*What is the relationship between self-efficacy and performance on national tests in mathematics in grade 8, and self-efficacy and performance on national tests in mathematics in grade 9?*	2.1	a)	Two proposed models (a latent levels model and a latent facets model) for students' self-efficacy resulted in excellent fit to the data
			b)	National test scores predicted students' self-efficacy a year later (except easy level self-efficacy), while their self-efficacy constructs in the first year did not predict national test scores in year two.
			c)	Students' self-efficacy was related to their national test scores both years
			d)	Students' performance scores were three to four times as stable from year one to year two, compared with their levels and facets of self-efficacy
RQ 2:	*Do the relationships between self-efficacy and performance on national tests differ according to levels or facets of self-efficacy?*	2.2	a)	The strongest effects were between performance scores and self-efficacy for medium difficulty and hard level, and performance scores and self-efficacy for perseverance (not giving up)
			b)	Students' self-efficacy for medium difficulty and hard level were predicted by their self-efficacy the first year, while the stability of self-efficacy for easy level was not significant from one year to the other
		* Finding not tested for statistical significance		

Research questions for each study:		Findings from each study:		
	Study 3: Stability and change of students' self-efficacy when learning a new topic			
RQ 1:	*What is the structural validity of the proposed self-efficacy constructs, which include facets and levels of self-efficacy?*	3.1	a)	It was reasonable to measure students' mathematics self-efficacy with a multidimensional instrument including latent constructs representing levels of difficulty and facets for classroom learning
RQ 2:	*How stable is the rank order of students' mathematics self-efficacy across a sequence of classroom lessons in mathematics?*	3.2	a)	The simplex model for self-efficacy change provided a good fit to my data
			b)	The stability of the rank order of students' self-efficacy was high when learning a new topic
			c)	The stability of the rank order of students' self-efficacy did not change significantly across the lesson sequence
			d)	The stability of the rank order of students' self-efficacy appeared similar across different levels of task difficulty
RQ 3:	*How does students' mathematics self-efficacy change across a sequence of classroom lessons in mathematics?*	3.3	a)	The mean-level of students' self-efficacy changed significantly across a sequence of lessons (except in the case of grade 10 students' self-efficacy for easy tasks)
			b)	The shape of change was best characterised by an initially steeper incline, followed by a flattening trend (e.g. non-linear growth) (except in the case of grade 6 students' self-efficacy for easy tasks)
			c)	For medium difficulty and hard tasks, grade 6 students' self-efficacy was significantly stronger at lesson 1, and increased more steeply, as compared with grade 10 students
			d)	The slope of students' self-efficacy was steeper in association with higher task difficulty *
		* Finding not tested for statistical significance		

Research questions for each study:		Findings from each study:		
	Study 4: The relationship between self-efficacy and mastery experiences across a sequence of lessons in mathematics			
RQ 1:	*What is the structural validity of the proposed mastery experiences constructs, which include facets and levels of mastery experiences?*	4.1	a)	It was reasonable to measure students' mathematics mastery experiences with a multidimensional instrument including latent constructs representing levels of difficulty and facets for classroom learning
RQ 2:	*What is the relationship between self-efficacy and mastery experiences across a sequence of regular lessons in mathematics?*	4.2	a)	Students' lesson self-efficacy was significantly more stable than their mastery experiences
			b)	The stability and cross-lagged paths of lesson self-efficacy and mastery experiences did not change over time
			c)	A reciprocal effects model was supported, where stronger self-efficacy predicted larger increases in mastery experiences, and more successful mastery experiences predicted larger gains in self-efficacy
			d)	The dominant cross-lagged effect was from lesson self-efficacy to mastery experiences (difference was significant)
			e)	Students' self-efficacy was most stable in relation to hard tasks, followed by medium difficulty and easy tasks
			f)	The stability of mastery experiences was weakest in relation to medium difficulty tasks
			g)	The cross-lagged effects between self-efficacy and mastery experiences were strongest in relation to medium difficulty tasks
RQ 3:	*What is the relationship between students' self-efficacy, mastery experiences, and ability conceptions?*	4.3	a)	Significant relationships between students' fixed beliefs in ability and their self-efficacy / mastery experiences supported the validity of my proposed measures
			b)	Higher fixed ability beliefs were related to weaker self-efficacy and mastery experiences
			c)	In regard to both mathematics self-efficacy and mastery experiences, the relationship was stronger with students' fixed beliefs in mathematics ability than their fixed beliefs in general ability *
			d)	There were significant relationships between students' ability conceptions and self-efficacy in a cross-lagged path model including also students' mastery experiences
			e)	The relationships between students' mastery experiences / self-efficacy and their ability conceptions differed according to levels of perceived difficulty, where stronger effects were found in relation to easier level tasks *
		* Finding not tested for statistical significance		

NB! In this table "more than" is a reflection of a relationship to other variables. In contrast, I have used "most" in an absolute sense, e.g. "most" is not necessarily "more than"

6.1 Cross-study findings and implications

When I look at my findings across each of the studies, I am able to address some theoretical themes of a more general nature that I could not adequately address within each study. I identified three key cross-study findings, which I will address next. These are based on findings that are supported across my studies, which serve to address one of the main themes of my thesis. For each of the cross-study findings, I first propose a theoretical position, before I discuss the findings from each study that gave rise to this position, in light of theory and previous research. Furthermore, I discuss implications as well as suggestions for future research in light of the main themes of the thesis. To recap, the main themes of my thesis, as set out in section 1.3, are a) the conceptualisation of self-efficacy, b) the relationship between self-efficacy and performance, and c) the development and change of self-efficacy. The three cross-study findings I identified are *1) Students consider levels of difficulty when assessing their mathematics performance experiences, and when forming their subsequent self-efficacy, 2) Higher accuracy but also quicker adjustment of self-efficacy with age, and 3) Novel classroom experiences predict mean-level changes in students' self-efficacy, but the novelty wears off*.

6.1.1 The conceptualisation of self-efficacy (and mastery experiences)

As set out in section 1.3.1, level of difficulty is an important part of the conceptualisation of the self-efficacy construct (Bandura, 1997; Zimmerman, 2000), but only few studies have taken level of difficulty into account when investigating self-efficacy in mathematics.

Cross-study finding 1: Students consider levels of difficulty when assessing their mathematics performance experiences, and when forming their subsequent self-efficacy

My findings demonstrated it is reasonable to measure students' mathematics self-efficacy and mastery experiences with multidimensional instruments including latent constructs representing levels of difficulty and test taking facets. In studies 1, 2, and 3, my findings provided support for a multidimensional model of students' self-efficacy, including latent levels of difficulty (findings 1.1, a; 2.1, a; and 3.1, a). In study 4 these findings were extended to include also support for a multidimensional model for students' mastery experiences, matched item-by-item with the measure for self-efficacy introduced and tested in study 3 (finding 4.1, a). In terms of further validity of my proposed measures, my findings in studies 1 and 2 demonstrated students' self-efficacy for easy, medium difficulty, and hard tasks were related to their performance on a national test (findings 1.2, a; 1.3, a and 2.1, c), while my findings in study 4 demonstrated expected relationships between students' ability conceptions (as measured by an established measure) and their self-efficacy and mastery experiences (finding 4.3, a). Overall, this provides support for the proposed measures, and gives compelling evidence that students consider levels of difficulty when forming their self-efficacy (for taking a test and for classroom learning), and when assessing their mastery experiences.

If students consider levels of difficulty when assessing their experiences and forming subsequent self-efficacy, this implies they would benefit from developing their skills in assessing task difficulty, choosing adequately challenging tasks, and appraising their performances in relation to the difficulty of the task. It is likely these skills can be developed in the classroom, with teacher support. In order to support positive self-efficacy that reflect students' capabilities on different levels of difficulty tasks, students might benefit from experiences with tasks of different levels of difficulty, and teacher scaffolding and feedback

to support both positive performance experiences and adaptive interpretation of their capabilities relative to the task requirements.

Each of the studies of the thesis provided evidence for the benefit of including levels of perceived difficulty when investigating students' mathematics self-efficacy in terms of substantive findings. In studies 1 and 2 I highlighted differences according to task difficulty in the strength of the relationship (e.g. accuracy) between students' self-efficacy and their performance scores (findings 1.2, c; 1.3, c; 2.2, a), and the stability of self-efficacy (2.2, b). In study 3 I found lower intercept and steeper growth of students' mean-level self-efficacy for harder tasks (finding 3.3, d), while in study 4 I found indications that the strength of the relationship between students' self-efficacy / mastery experiences and their ability conceptions varied as an effect of task difficulty (finding 4.3, e). These findings were not tested for statistical significance, as the level constructs were generally specified in separate models. It would be valuable in future research to further investigate the role of perceived task difficulty in relation to students' self-efficacy, considering themes such as stability, mean-level changes, relationship with performance / mastery, and the relationship with other constructs such as ability conceptions.

6.1.2 The relationship between self-efficacy and performance

Previous research has asserted self-beliefs tend to become more differentiated, negative and more closely associated to achievement with age (Midgley et al., 1989; Valentine et al., 2004). Similarly, research has indicated the strength of the relationship between self-efficacy and performance experiences (sometimes termed accuracy) is likely to vary according to age, where older students are better at assessing their own performance capabilities (Bong, 2009; Multon et al., 1991; see also section 1.3.2). In line with this, the effect of self-efficacy on

performance has been empirically supported in the case of adults, but not children (Talsma et al., 2018).

Cross-study finding 2: Higher accuracy but also quicker adjustment of self-efficacy with age

In support of theory and previous research, in study 1 of my thesis (findings 1.2, b and 1.3, b) I found empirical support for a stronger relationship between self-efficacy and performance as an effect of age, in study 2 (finding 2.1, b) I found significant effects from performance to students' self-efficacy but not in the other direction, and in study 3 I found older students' self-efficacy was significantly weaker (e.g. less positive) as compared with the younger students (finding 3.3, d). Furthermore, in study 3 (finding 3.3, c), I found significantly flatter growth shape of grade 10 as compared with grade 6 students, which might reflect that the older students adjusted their self-efficacy more quickly than the younger students (after having been introduced to a new topic). Some researchers have used the term *equilibrium* to refer to accurate self-efficacy. Williams & Williams (2010) argued 15-year old students (e.g. grade 10 in Norway) had reached equilibrium in terms of developing realistic self-appraisals in mathematics, through their classroom experiences over 8-10 years.

In addition to providing support for previous findings (e.g. Bong, 2009; Multon et al., 1991) that the strength of the relationship between self-efficacy and performance experiences increases as an effect of cognitive maturity / better self-assessment skills (e.g. Nicholls, 1978; Stipek, 1981), and that older students' self-efficacy is more negative as compared to younger students' (e.g. Bong, 2009), my findings also indicate that older students adjust their self-efficacy faster, in the context of performance experiences with a new topic. As there was no objective performance measure in study 3, it is uncertain whether the older students' weaker and more "stable" self-efficacy reflected their capabilities.

The flatter growth of older compared with younger students' self-efficacy when learning a new topic might reflect more stable self-perceptions (and cognitive appraisals, see section 4.5.3) in terms of classroom rank-order standing. While students are increasingly capable of making ability comparisons with age / cognitive maturity (e.g. Nichols, 1978), researchers have argued the school environment for older students also directs students' attention away from a mastery /improvement focus, toward a focus on ability relative to others (Ames & Ames, 1984; Anderman & Midgley, 1997). While focusing on mastery and individual learning gains likely is beneficial to all students (as all students have the possibility to make progress in their own learning), only few students (the highest ranked) can be "successful" in comparative structures. If students' self-perceptions about ability in mathematics become more stable with age, early intervention is crucial to ensure the kinds of appraisals students use when assessing their self-efficacy building information is adaptive. Students might benefit from teacher modelling of adaptive appraisal, for instance how to base assessment on task information, rather than bias /previous perceptions.

It is not certain the more accurate / realistic self-efficacy of older students are beneficial to them, in terms of learning or motivation gains (Klassen, 2004). On the contrary, theory posits that optimistic self-efficacy predict e.g. higher effort, increased persistence, and gains in performance (Bandura, 1997). While older students' self-efficacy might be more accurate, this accuracy might be in relation to their rank-order standing in the classroom. If information about relative ability is both salient and available to older students, they might be losing out on valuable self-efficacy building information from their own learning gains (mastery experiences). If students' self-efficacy declines with age partially because of the learning environment, this has important implications for practice. Teachers might try to shift the focus

in the classroom away from relative ability, and support students' skills at gauging their own learning progress.

While research has indicated higher *accuracy* of self-efficacy with age, we need to increase our knowledge about whether, and if so, how the effects between self-efficacy and learning and performance *gains* differ according to age. If younger students' self-efficacy is less stable (and more amenable to change) than older students, the gains from performance experiences on self-efficacy might be larger for younger children. In contrast, older students might benefit (in terms of performance and learning gains) more from increased self-efficacy than younger students, due to better self-regulation skills. Although Talsma et al. (2018) demonstrated the effect from children's self-efficacy on their performance was non-significant, only two studies including school-age children was included in the meta-analysis. Contrary to this, in study 4 (finding 4.2, c) I demonstrated that grade 6 and grade 10 students' self-efficacy significantly predicted gains in their mastery experiences. However, in my study I did not investigate whether this effect differed according to age, and I used a measure of mastery experiences, rather than objective performance. In future studies it will be valuable to increase our knowledge about the theoretically proposed *benefits* from stronger self-efficacy, through investigating the reciprocal effects between self-efficacy and performance experiences in mathematics, and as well as factors that might moderate these effects (e.g. age, gender, task difficulty, and cognitive appraisals).

6.1.3 The development and change of self-efficacy

Theoretical tenets posit self-efficacy changes as an effect of information from influential events, through four sources of self-efficacy that are also cognitively appraised (Bandura, 1997; Usher & Pajares, 2008). Contrary to research within for instance management theory, a reciprocal effects model for self-efficacy in mathematics appears not to have been

investigated within a process framework: e.g. the reciprocal relationship between self-efficacy and performance experiences across a sequence of events (e.g. Shea & Howell, 2000). The scarcity of studies that investigated reciprocal relationships in terms of mathematics and with school age children, and the fact most studies have used measures of performance, not mastery, across long measurement gaps, means the theoretically proposed reciprocal effects model across influential events, reflecting the process of self-efficacy change, stands largely untested in relation to school-age children's mathematics self-efficacy and mastery experiences.

Cross-study finding 3: Novel classroom experiences predict mean-level changes in students' self-efficacy, but the novelty wears off

In study 2 my findings replicated previous studies, in that the cross-lagged effect from performance on self-efficacy, but not self-efficacy on performance, was significant across the course of one year (finding 2.1, b). In contrast to this long measurement gap, the design of phase 2 enabled investigation into the *process* of self-efficacy change. In study 3 I found evidence that students' self-efficacy changed as an effect of the most recent self-efficacy building experiences (e.g. the simplex model of change, finding 3.2, a), and that over a sequence of lessons when introduced to a new topic, the mean-level of students' self-efficacy grew significantly (finding 3.3, a). The growth was steepest in association with hard tasks (finding 3.3, d). In study 4 I demonstrated in cross-lagged path models that stronger self-efficacy predicted gains in mastery experiences, and stronger mastery experiences predicted gains in self-efficacy (finding 4.2, c). Considering studies 3 and 4 used the same data (e.g. the same students across the same learning events), this indicates that positive student mastery experiences when engaging with a new topic were related to mean-level increases in students' self-efficacy. In other words, the effect from mastery experiences on gains in self-efficacy

(study 4, finding 4.2, c) "drove" the mean-level changes in self-efficacy (study 3, finding 3.3, a). Furthermore, the growth of self-efficacy flattened after a couple of lessons (finding 3.3, b), while the effect of mastery experiences on gains in self-efficacy remained stable across the whole sequence (finding 4.2, b), indicating students experienced more / stronger mastery experiences early on in the sequence of lessons.

I argue it is likely that the first performance experiences when engaging with a new topic were highly salient to students in terms of self-efficacy building information, and that the flattening growth reflected a tendency toward equilibrium between students' self-efficacy and mastery experiences, as students gained experience with the new topic. Overall, my findings provide support for the position that students' self-efficacy does not change mainly as an effect of time passing, nor as an effect of an "accumulation" of performance experiences, but rather as an effect of the saliency of each learning event. It is important to consider the context within which these self-efficacy changes were produced: students were during the first lesson introduced to a new topic in mathematics. It is possible the novelty of the topic is crucial to students' mastery experiences. Stylianides and Stylianides (2014) demonstrated the memorability of tasks was important to facilitate changes in students' beliefs about mathematics. Furthermore, the participants in the study by Stylianides and Stylianides (2014) highlighted the difficulty of the task they engaged with, indicating that overcoming a perceived difficult challenge on a memorable task was important.

In section 4.5.4 I argued that students might benefit more also from their mastery experiences with non-novel tasks if the teacher highlights the differences to previous tasks in terms of students' learning gains. Also, while teachers cannot provide "intervention-style" instruction all the time, they can make efforts that all students occasionally engage with memorable and

challenging tasks, which with the benefit of teacher support are likely to provide salient sources for the students' mathematics self-efficacy. Considering the scarce evidence available regarding the cross-lagged (reciprocal) relationships between students' mathematics self-efficacy and performance, these findings are an important step toward increasing our knowledge about the process of self-efficacy change in a classroom setting. Future studies are needed to better understand the interaction between changes in students' self-efficacy in relation to factors such as classroom quality and climate, the tasks students engage with, and how students appraise of their experiences with mathematics. It would be valuable to also include data from different sources, such as teacher and student interviews and observation data, in order to triangulate findings, and gain better insight into how factors at the individual (e.g. appraisal) and group (e.g. teaching quality) levels interact in terms of influences on students' self-efficacy.

6.2 Limitations

While I discussed limitations of each of my studies previously, I will here emphasise some important limitations that apply to the thesis as a whole. First, I will consider my sample. For the present thesis I did not include a random sample, rather my participants were purposively collected. While I made efforts to ensure generalizability (see sections 2.3.1 and 4.3.2) it is possible my sample was not representative to the Norwegian school student population (grades 5-10). As per the discussion in section 4.5.4, the growth of students' self-efficacy may differ as an effect of socio-economic background. While the Norwegian education system provides an advantageous setting with little stratification (see section 2.3.1), I could not test whether my sample was representative, as I did not have access to background data at the individual level. Future replications of my studies with different samples might provide new insights.

Second, for different research question within my thesis, I alternated between merging the different grade levels included in my samples, and investigating them separately. There are advantages and disadvantages to both approaches – mainly, if the effects I am investigating differ for different students of different grades (age), including them in the same sample might mean group-level variation is "hidden". At the same time, if the samples are investigated in separate models, differences according to age cannot be statistically tested. In terms of the conceptualisation of self-efficacy, my investigations in study 1 indicated there were no differences to the measurement structure of self-efficacy due to age. However, in study 3 I found indications the meaning of one item used to measure self-efficacy changed meaning across the lessons for grade 6 students (see section 4.4.2.3), and similarly in study 4 the meaning of two items used to measure mastery experiences might have changed meaning, also in regard to grade 6 students (see section 5.4.1.3). This might reflect the younger students' growing conceptualisations of their own confidence and capability in regard to mathematics. Accordingly, any inferences across grades about the conceptualisation of self-efficacy as well as the relationship between self-efficacy and other constructs should be made with caution. There are models that allow for differentiation across groups, such as the grade-as-moderator groups I specified in study 3 (see section 4.4.4.2) in order to compare grade 6 and 10 intercept and slope, while allowing different growth trajectories for each group. It would be valuable to ascertain in future studies to what degree student age or grade level moderate theoretical models of self-efficacy.

As an extension of my second point, while I conceptualised self-efficacy as an individual construct, most of my analyses were at the group level. As per the concept of ergodicity, it is important not to make individual assumptions based on group-level trends (Molenaar, 2004). In terms of students' self-efficacy; the growth, stability, and reciprocal effects at the

individual level might differ to my group-level findings. In study 3 I used latent growth models to investigate the growth of students' self-efficacy across learning events, and found that in the case of grade 6 students there was a significant variance around the mean slope parameter, indicating there were individual trajectories for the growth of students' self-efficacy. In future research diary studies (e.g. Schmitz, 2006) might illuminate how the theoretically proposed models of self-efficacy stability and change, and the relationship with mastery experiences, differ across students.

One of the recurring themes in this thesis is the focus on students' perceptions – students' perceptions of their own capability to carry out future tasks, their perceptions of past performance events, and their perceptions of task difficulty. An advantage to this approach is we can investigate whether there are trends across students in terms of their perceptions. On the other hand, a disadvantage is it is challenging to "translate" such trends into classroom practice. For example, I have included students' perceived levels of difficulty rather than objective task difficulty. While I argue it is reasonable to investigate perceived level of difficulty, given that the same task may present different challenges for different children, a recommendation such as "Students might benefit from working on medium difficulty tasks, in relation to developing adaptive self-efficacy" does not provide teachers with information about task types. In contrast, an important implication from my thesis is that teachers might need to ascertain not only the "actual" strength and weaknesses of all children, in terms of their skills in mathematics, but also how students perceive of and make sense of different tasks and their level of difficulty. In future studies it would be valuable to measure students' self-efficacy for different levels of mathematical tasks through briefly displaying the tasks, similar to the approach often used for self-efficacy for test performance (e.g. Dowling, 1978), and in addition ask student to rate how difficult they perceive each task to be. These measures

could then be matched with students' mastery experiences from the same tasks, as per my approach in study 4. This could expand our knowledge regarding self-efficacy and mastery experiences according to actual as well as perceived task difficulty.

Finally, while in my studies I discuss ways in which teachers might work to support students' adaptive self-efficacy, I did not include any measures of teachers' current practices. Thus, any suggestions for teacher practice are not intended as judgment calls of what teachers do or do not do, but rather as a theoretical discussion regarding what might be advantageous to do (or perhaps do more of), considering the findings from my thesis in relation to previous theory and research.

6.3 Conclusion

In this thesis I set out to investigate a series of research questions, framed within separate studies, with an aim to increase our knowledge of an overarching theme: students' self-efficacy, and the relationship between students' self-efficacy and their experiences with mathematics. Through investigating these substantive themes, I have also developed methodologies for investigating multidimensional self-efficacy and mastery experiences. I will briefly describe how each of my four studies contributed to my overarching aim, before I make some concluding remarks on the role of the teacher in light of the findings from this thesis.

Study 1

In study 1, I investigated the structural validity of a new measure for self-efficacy, which disentangles test taking facets and levels of difficulty from strength of self-efficacy. My findings indicated that students consider levels of perceived difficulty when forming their self-efficacy, and the measure was validated through significant relationships with students'

test performance. Study 1 provided an important baseline for the further investigations of my thesis, both in terms of methodology (the proposed measurement model) and substantive focus (the conceptualisation of self-efficacy, and investigating differential effects according to levels of perceived difficulty).

Study 2

Study 2 built on the measurement model and substantive themes from study 1, and the findings demonstrated that over the course of one year performance scores predicted subsequent mathematics self-efficacy, but not vice versa. This study was of a smaller scope as compared with the other studies in my thesis, but made an important empirical contribution considering the scarcity of investigations into reciprocal effects in mathematics and with school-age children.

Study 3

In study 3 I used descriptive statistics, autoregressive paths models, and latent growth models for change as complimentary approaches to investigate the stability and change of students' self-efficacy across a sequence of lessons in mathematics when they were introduced to a new topic. The proposed measurement model (adapted for a classroom learning context) was supported, and differential effects according to level of perceived difficulty justified the substantive focus set forth in study 1. While study 1 prepared the ground for all three subsequent studies, my third study built on and extended the findings from studies 1 and 2, while also introducing a new dimension to my overall investigation, in terms of the classroom learning context and a micro-analytic approach.

The autoregressive path models demonstrated high rank-order stability of students' self-efficacy, even as students were introduced to new topics in mathematics. Furthermore, latent growth models demonstrated significant mean-level changes to students' self-efficacy across the sequence of lessons. I know of no previous research investigating mean-level changes in students' mathematics self-efficacy in a micro-analytic framework, e.g. across a sequence of learning events over a short period of time. In view of the lack of previous findings to serve as comparison, study 3 contributed with important, novel, information about significant mean-level changes in students' self-efficacy across a sequence of regular lessons when students were introduced to a new mathematical topic, even as their rank-order standing remained stable. The findings from study 3 demonstrated differences in mean-level growth according to both age and task difficulty, and are a first step to increasing our knowledge about the process of mathematics self-efficacy change in a classroom setting.

Study 4

As the final study of my thesis, study 4 addressed some important questions emerging from studies 1 to 3; provided additional support for the measurement model including levels of difficulty; and further demonstrated the benefits of including levels of perceived difficulty in substantive investigations about students' self-efficacy in mathematics. In studies 1, 2, and 3, my findings provided support for a multidimensional model of students' self-efficacy, including latent levels of difficulty. In study 4 these findings were extended to include also support for a multidimensional model for students' mastery experiences, along with a methodology for investigating reciprocal effects using matched measures of self-efficacy and mastery experiences. Overall, this gives compelling evidence that students consider levels of difficulty when assessing their performance experiences and forming their self-efficacy in mathematics. My findings in study 4 provided empirical support to the validity of the

measures of self-efficacy as well as mastery experiences, through demonstrating significant relationships with students' ability conceptions. Furthermore, the benefits from including levels of perceived task difficulty in substantive investigations were further highlighted through indications the relationship between ability conceptions and self-efficacy and mastery experiences is progressively weaker with higher task difficulty.

The new measure of mastery experiences proposed in study 4 enabled investigation into the theoretically proposed process of self-efficacy expectation change, including cycles of mastery experiences and self-efficacy. In studies 1 and 2 I included measures of students' (national test) performance, however performance was in these studies treated as a dependent (and independent, in the case of study 2) variable, and not measured sequentially to students' self-efficacy. This stands in contrast with the strategic design of study 4 which enabled me to treat mastery experiences as a key ingredient in the process of self-efficacy change when students were introduced to new topics. This investigation contributed important information about the relative stability of the constructs (lesson self-efficacy more stable than mastery experiences), as well as the dominant effect (from self-efficacy to mastery experiences).

Overall, through empirical investigations within each of the four studies this thesis has made significant contributions to our knowledge of students' self-efficacy and experiences with mathematics. While I have discussed several implications stemming from the findings within this thesis, one major implication for schools and teachers is that in order to support student engagement and learning with mathematics, we need to focus on students' appraisals as well as their performance. There is a growing consensus among researchers about the importance of affective aspects of learning mathematics. Similar to this, I argue that supporting adaptive sense-making of students' experiences should not be viewed as separate to teaching, but

rather as an integral part of supporting student learning and development. While this is not a controversial statement, it might be an important one. Teachers have limited time and resources available, and shifting the focus from *what is taught*, beyond even *what is learned*, to arrive at *what meaning is made*, likely requires adjustments to many teachers' practices as well as to teacher education.

References

Ames, C., & Ames, R. (1984). Systems of student and teacher motivation: Toward a qualitative definition. *Journal of Educational Psychology*, *76*(4), 535–556. http://doi.org/10.1037//0022-0663.76.4.535

Anderman, E. M., & Midgley, C. (1997). Changes in achievement goal orientations, perceived academic competence, and grades across the transition to middle-level schools. *Contemporary Educational Psychology*, *22*(3), 269–298. http://doi.org/10.1006/ceps.1996.0926

Ayotola, A., & Adedeji, T. (2009). The relationship between gender, age, mental ability, anxiety, mathematics self-efficacy and achievement in mathematics. *Cypriot Journal of Educational Sciences*, *4*, 113–124. Retrieved from http://www.world-education-center.org/index.php/cjes/article/viewArticle/110

Bandalos, D. L. (2002). The Effects of Item Parceling on Goodness-of-Fit and Parameter Estimate Bias in Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(1), 78–102. http://doi.org/10.1207/S15328007SEM0901_5

Bandalos, D. L., & Finney, S. J. (2001). Item Parceling Issues in Structural Equation Modelin. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New Developments and Techniques in Structural Equation Modeling* (pp. 269–296). Taylor & Francis e-library.

Bandura, A. (1977). Self-efficacy: toward a unifying theory of behavioral change. *Psychological Review*, *84*(2), 191–215. http://doi.org/10.1037/0033-295X.84.2.191

Bandura, A. (1986). *Social foundations of thought and action: a social cognitive theory*. Englewood Cliffs, N.J.: Prentice-Hall.

Bandura, A. (1989). Human agency in social cognitive theory. *American Psychologist*, *44*(9), 1175–84. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2782727

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: W. H. Freeman and Co.

Bandura, A. (2006). Guide for constructing self-efficacy scales. In F. Pajares & T. C. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (pp. 307–337). Greenwich: Information Age Publishing.

Bandura, A. (2012). On the Functional Properties of Perceived Self-Efficacy Revisited. *Journal of Management*, *38*(1), 9–44. http://doi.org/10.1177/0149206311410606

Bandura, A., & Schunk, D. H. (1981). Cultivating competence, self-efficacy, and intrinsic interest through proximal self-motivation. *Journal of Personality and Social Psychology*, *41*(3), 586–598. http://doi.org/10.1037/0022-3514.41.3.586

Bandura, A., & Wood, R. (1989). Effect of perceived controllability and performance standards on self-regulation of complex decision making. *Journal of Personality and Social Psychology*, *56*(5), 805–814. http://doi.org/10.1037/0022-3514.56.5.805

Benenson, J. F., & Dweck, C. S. (1986). The Development of Trait Explanations and Self-Evaluations in the Academic and Social Domains. *Child Development*, *57*(5), 1179–1187.

Bennett, N., & Desforges, C. (1988). Matching Classroom Tasks to Students' Attainments. *The Elementary School Journal*, *88*(3), 220–234.

Bernacki, M. L., Nokes-Malach, T. J., & Aleven, V. (2015). Examining self-efficacy during learning: variability and relations to behavior, performance, and learning. *Metacognition and Learning*, *10*(1). http://doi.org/10.1007/s11409-014-9127-x

Berry, J. M., West, R. L., & Dennehey, D. M. (1989). Reliability and validity of the Memory Self-Efficacy Questionnaire. *Developmental Psychology*, *25*(5), 701–713.

http://doi.org/10.1037/0012-1649.25.5.701
Betz, N. E., & Hackett, G. (1983). The relationship of mathematics self-efficacy expectations to the selection of science-based college majors. *Journal of Vocational Behavior*, *23*(3), 329–345. http://doi.org/10.1016/0001-8791(83)90046-5
Bollen, K. A., & Long, J. S. (1983). *Testing structural equation models*. (K. A. Bollen & J. S. Long, Eds.). Newbury Park, CA: Sage.
Bong, M. (2001). Between- and within-domain relations of academic motivation among middle and high school students: Self-efficacy, task-value, and achievement goals. *Journal of Educational Psychology*, *93*(1), 23–34.
Bong, M. (2002). Predictive utility of subject-, task-, and problem-specific self-efficacy judgments for immediate and delayed academic performances. *The Journal of Experimental Education*, *70*(2), 133–162.
Bong, M. (2005). Within-grade changes in Korean girls' motivation and perceptions of the learning environment across domains and achievement levels. *Journal of Educational Psychology*, *97*(4), 656–672. http://doi.org/10.1037/0022-0663.97.4.656
Bong, M. (2006). Asking the right question: How confident are you that you could successfully perform these tasks ? In F. Pajares & T. C. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (pp. 287–305). Greenwich: Information Age Publishing.
Bong, M. (2009). Age-related differences in achievement goal differentiation. *Journal of Educational Psychology*, *101*(4), 879–896. Retrieved from http://search.proquest.com/docview/614546142?accountid=38978
Bong, M., & Hocevar, D. (2002). Measuring Self-Efficacy: Multitrait-Multimethod Comparison of Scaling Procedures. *Applied Measurement in Education*, *15*(2), 143–171. http://doi.org/10.1207/S15324818AME1502
Bong, M., & Skaalvik, E. M. (2003). Academic self-concept and self-efficacy: How different are they really? *Educational Psychology Review*, *15*, 1–40.
Bonne, L. (2012). *The Effects of Primary Students` Mathematics Self-efficacy and Beliefs about Intelligence on Their Mathematics Achievement : A Mixed-methods Intervention Study*. Victoria University of Wellington.
Bonne, L., & Johnston, M. (2016). Students' beliefs about themselves as mathematics learners. *Thinking Skills and Creativity*, *20*, 17–28. http://doi.org/10.1016/j.tsc.2016.02.001
Bråten, I., Samuelstuen, M. S., & Strømsø, H. I. (2004). Do students' self-efficacy beliefs moderate the effects of performance goals on self-regulatory strategy use? *Educational Psychology*, *24*(2), 231–247. http://doi.org/10.1080/0144341032000160164
Bråten, I., & Strømsø, H. I. (2005). The relationship between epistemological beliefs, implicit theories of intelligence, and self-regulated learning among Norwegian postsecondary students. *The British Journal of Educational Psychology*, *75*(Pt 4), 539–65. http://doi.org/10.1348/000709905X25067
Britner, S. L., & Pajares, F. (2006). Sources of science self-efficacy beliefs of middle school students. *Journal of Research in Science Teaching*, *43*(5), 485–499. http://doi.org/10.1002/tea.20131
Buehl, M. M., & Alexander, P. A. (2001). Beliefs About Academic Knowledge. *Educational Psychology Review*, *13*(4), 385–418.
Bussey, K., & Bandura, A. (1999). Social cognitive theory of gender development and differentiation. *Psychological Review*, *106*(4), 676–713.
Butler, R. (2000). Making judgments about ability: The role of implicit theories of ability

in moderating inferences from temporal and social comparison information. *Journal of Personality and Social Psychology*, *78*(5), 965–978.
Butz, A. R., & Usher, E. L. (2015). Salient sources of early adolescents' self-efficacy in two domains. *Contemporary Educational Psychology*, *42*, 49–61. http://doi.org/10.1016/j.cedpsych.2015.04.001
Byars-Winston, A., Diestelmann, J., Savoy, J. N., & Hoyt, W. T. (2017). Unique effects and moderators of effects of sources on self-efficacy: A model-based meta-analysis. *Journal of Counseling Psychology*, *64*(6), 645–658. http://doi.org/10.1037/cou0000219
Cairney, J., Hay, J. A., Faught, B. E., Leger, L., & Mathers, B. (2008). Generalized self-efficacy and performance on the 20-metre shuttle run in children. *American Journal of Human Biology : The Official Journal of the Human Biology Council*, *20*(2), 132–138. http://doi.org/10.1002/ajhb.20690
Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105.
Caprara, G. V., Fida, R., Vecchione, M., Del Bove, G., Vecchio, G. M., Barbaranelli, C., & Bandura, A. (2008). Longitudinal analysis of the role of perceived self-efficacy for self-regulated learning in academic continuance and achievement. *Journal of Educational Psychology*, *100*(3), 525–534. Retrieved from http://search.proquest.com/docview/614483692?accountid=38978
Caprara, G. V., Vecchione, M., Alessandri, G., Gerbino, M., & Barbaranelli, C. (2011). The contribution of personality traits and self-efficacy beliefs to academic achievement: A longitudinal study. *British Journal of Educational Psychology*, *81*(1), 78–96. http://doi.org/10.1348/2044-8279.002004
Chen, F., Bollen, K. A., Paxton, P., Curran, P. J., & Kirby, J. B. (2001). Improper solutions in Structural equation models. *Sociological Methods & Research*, *29*(4), 468–508.
Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464–504. http://doi.org/10.1080/10705510701301834
Chen, J. A., & Pajares, F. (2010). Implicit theories of ability of Grade 6 science students: Relation to epistemological beliefs and academic motivation and achievement in science. *Contemporary Educational Psychology*, *35*(1), 75–87. http://doi.org/10.1016/j.cedpsych.2009.10.003
Chen, P., & Zimmerman, B. J. (2007). A cross-national comparison study on the accuracy of self-efficacy beliefs of middle-school mathematics students. *Journal of Experimental Education*, *75*(3), 221–244. http://doi.org/10.3200/JEXE.75.3.221-244
Chiorri, C. (2014). Competing Factor Structures for the Toronto Empathy Questionnaire. In D. Wyatt & J. Panksepp (Eds.), *Psychology of empathy* (pp. 1–61). New York: Nova Science Publishers.
Christensen, L., & Mendoza, J. L. (1986). A Method of Assessing Change in a Single Subject : An Alteration of the RC Index. *Behavior Therapy*, *17*, 305–308.
Cooper, H. M. (1979). Pygmalion Grows Up : A Model for Teacher Expectation Communication and Performance Influence. *Review of Educational Research*, *49*(3), 389–410.
Davis-Kean, P. E., Huesmann, L. R., Jager, J., Bates, J. E., Collins, W. A., & Lansford, J. E. (2008). Changes in the Relation of Self-Efficacy Beliefs and Behaviors across Development. *Child Development*, *79*(5), 1257–1269.
Davis, J. L., Burnette, J. L., Allison, S. T., & Stone, H. (2011). Against the odds: academic

underdogs benefit from incremental theories. *Social Psychology of Education*, *14*(3), 331–346. http://doi.org/10.1007/s11218-010-9147-6
Dierdorff, E. C., Surface, E. A., & Brown, K. G. (2010). Frame-of-reference training effectiveness: effects of goal orientation and self-efficacy on affective, cognitive, skill-based, and transfer outcomes. *The Journal of Applied Psychology*, *95*(6), 1181–1191. http://doi.org/10.1037/a0020856
Dillon, W. R., Kumar, A., & Mulani, N. (1987). Offending estimates in covariance structure analysis: Comments on the causes of and solutions to Heywood cases. *Psychological Bulletin*, *101*(1), 126–135. http://doi.org/10.1037/0033-2909.101.1.126
Dowling, D. M. (1978). *The development of a mathematics confidence scale and its application in the study of confidence in women college students*. Ohio State University.
Driver, C. C., & Voelkle, M. C. (2016). Hierarchical Bayesian Continuous Time Dynamic Modeling, (November).
Dweck, C. S. (2000). *Self-theories: Their role in motivation, personality, and development*. Philadelphia, PA: Psychology Press.
Dweck, C. S. (2008). *Self-theories: Their role in motivation, personality and development*. San Diego: Academic Press.
Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review*, *95*(2), 256–273. http://doi.org/10.1037//0033-295X.95.2.256
Ellingsæter, A. L., & Leira, A. (2006). *Politicising parenthood in Scandinavia*. Bristol, UK: The Policy Press.
Fast, L. A., Lewis, J. L., Bryant, M. J., Bocian, K. a., Cardullo, R. a., Rettig, M., & Hammond, K. a. (2010). Does math self-efficacy mediate the effect of the perceived classroom environment on standardized math test performance? *Journal of Educational Psychology*, *102*(3), 729–740. http://doi.org/10.1037/a0018863
Fennema, E., & Sherman, J. A. (1976). Fennema-Sherman Mathematics Attitudes Scales : Instruments Designed to Measure Attitudes toward the Learning of Mathematics by Females and Males. *Journal for Research in Mathematics Education*, *7*(5), 324–326.
Friedel, J. M., Cortina, K. S., Turner, J. C., & Midgley, C. (2010). Changes in efficacy beliefs in mathematics across the transition to middle school: Examining the effects of perceived teacher and parent goal emphases. *Journal of Educational Psychology*, *102*(1), 102–114. Retrieved from http://search.proquest.com/docview/614511223?accountid=38978
Gamlem, S. M., & Munthe, E. (2014). Mapping the quality of feedback to support students' learning in lower secondary classrooms. *Cambridge Journal of Education*, *44*(1), 75–92. http://doi.org/10.1080/0305764X.2013.855171
Ganley, C. M., & Vasilyeva, M. (2014). The role of anxiety and working memory in gender differences in mathematics. *Journal of Educational Psychology*, *106*(1), 105–120. http://doi.org/10.1037/a0034099
Gerbing, D. W., & Anderson, J. C. (1987). Improper solutions in the analysis of covariance structures: their interpretability and a comparison of alternative respecifications. *Psychometrika*, *52*(1), 99–111.
Gibbs, B. G. (2010). Reversing fortunes or content change? Gender gaps in math-related skill throughout childhood. *Social Science Research*, *39*, 540–569.
Gist, M. E., Schwoerer, C., & Rosen, B. (1989). Effects of alternative training methods on self-efficacy and performance in computer software training. *Journal of Applied Psychology*, *74*(6), 884–891. http://doi.org/10.1037/0021-9010.74.6.884

Gottfried, A. E., Fleming, J. S., & Gottfried, A. W. (2001). Continuity of Academic Intrinsic Motivation From Childhood Through Late Adolescence: A Longitudinal Study. *Journal of Educational Psychology*, *93*(1), 3–13.

Graham, J. W., & Coffman, D. L. (2012). Structural equation modeling with missing data. In R. H. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (pp. 277–295). New York: Guildford press.

Hackett, G. (1995). Self-efficacy in careeer choice and development. In A. Bandura (Ed.), *Self-efficacy in changing societies* (pp. 232–258). New York: Cambridge University Press.

Hackett, G., & Betz, N. E. (1989). An exploration of the mathematics self-efficacy/ mathematics performance correspondence. *Journal for Research in Mathematics Education*, *20*(3), 261–273.

Hannula, M. S. (2012). Exploring new dimensions of mathematics-related affect: embodied and social theories. *Research in Mathematics Education*, *14*(2), 137–161. http://doi.org/10.1080/14794802.2012.694281

Hannula, M. S., Bofah, E. A., Tuohilampi, L., & Metsämuuronen, J. (2014). A longitudinal analysis of the relationship between mathematics-related affect and achievement in Finland. In *Proceedings of the Joint Meeting of PME 38 and PME-NA 36* (Vol. 38, pp. 249–256).

Hannula, M. S., Maijala, H., & Pehkonen, E. (2004). Development of Understanding and Self- Confidence in Mathematics ; Grades 5 – 8. In *Proceedings of the 28th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 3, pp. 17–24).

Helle, L., Laakkonen, E., Tuijula, T., & Vermunt, J. D. (2013). The developmental trajectory of perceived self-regulation, personal interest, and general achievement throughout high school: A longitudinal study. *British Journal of Educational Psychology*, *83*(2), 252–266. http://doi.org/10.1111/bjep.12014

Hirai, K., Suzuki, Y., Tsuneto, S., Ikenaga, M., Hosaka, T., & Kashiwagi, T. (2002). A structural model of the relationships among self-efficacy, psychological adjustment, and physical condition in Japanese advanced cancer patients. *Psycho-Oncology*, *11*(3), 221–229. http://doi.org/10.1002/pon.561

Hopfenbeck, T. N. (2011). Fra teoretiske modeller til klasseromspraksis: Hvordan fremme selvregulert læring? *Norsk Pedagogisk Tidsskrift*, *5*, 360–373.

Hornstra, L., van der Veen, I., Peetsma, T., & Volman, M. (2013). Developments in motivation and achievement during primary school: A longitudinal study on group-specific differences. *Learning and Individual Differences*, *23*, 195–204. http://doi.org/10.1016/j.lindif.2012.09.004

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. http://doi.org/10.1080/10705519909540118

Jacobs, J. E., Lanza, S., Osgood, D. W., Eccles, J. S., & Wigfield, A. (2002). Changes in Children's Self-Competence and Values: Gender and Domain Differences across Grades One through Twelve. *Child Development*, *73*(2), 509–527.

Jacobson, N. S., & Truax, P. (1991). Clinical Significance : A Statistical Approach to Denning Meaningful Change in Psychotherapy Research, *59*(1), 12–19.

Joët, G., Usher, E. L., & Bressoux, P. (2011). Sources of self-efficacy: An investigation of elementary school students in France. *Journal of Educational Psychology*, *103*(3), 649–663. http://doi.org/10.1037/a0024048

Johnson, M. L., Edwards, O. V, & Dai, T. (2014). Growth Trajectories of Task Value and Self-efficacy Across an Academic Semester. *Universal Journal of Educational Research*, *2*(1), 10–18. http://doi.org/10.13189/ujer.2014.020102

Jungert, T., & Andersson, U. (2013). Self-efficacy Beliefs in Mathematics, Native Language Literacy and Foreign Language Amongst Boys and Girls with and without Mathematic Difficulties. *Scandinavian Journal of Educational Research*, *57*(1), 1–15. http://doi.org/10.1080/00313831.2011.621140

Kıran, D., & Sungur, S. (2011). Middle School Students' Science Self-Efficacy and Its Sources: Examination of Gender Difference. *Journal of Science Education and Technology*, *21*(5), 619–630. http://doi.org/10.1007/s10956-011-9351-y

Klassen, R. M. (2004). Optimism and realism: A review of self-efficacy from a cross-cultural perspective. *International Journal of Psychology*, *39*(3), 205–230. http://doi.org/10.1080/00207590344000330

Klassen, R. M., Bong, M., Usher, E. L., Chong, W. H., Huan, V. S., Wong, I. Y. F., & Georgiou, T. (2009). Exploring the validity of a teachers' self-efficacy scale in five countries. *Contemporary Educational Psychology*, *34*(1), 67–76. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0361476X08000556

Klassen, R. M., & Chiu, M. M. (2010). Effects on teachers' self-efficacy and job satisfaction: Teacher gender, years of experience, and job stress. *Journal of Educational Psychology*, *102*(3), 741–756. Retrieved from http://search.proquest.com/docview/745195893?accountid=38978

Klassen, R. M., & Usher, E. L. (2010). Self-efficacy in educational settings: Recent research and emerging directions. In T. C. Urdan & S. A. Karabenick (Eds.), *Advances in Motivation and Achievement* (Vol. 16, pp. 1–33). Bingley: Emerald Group Publishing. http://doi.org/10.1108/S0749-7423(2010)000016A004

Klette, K. (2011). Summary on the conference "Lower secondary schools in Norway", 03.11.2011. In *Utdanning 2020*.

Klette, K., Lie, S., Ødegaard, M., Anmarkrud, Ø., Arnesen, N., Bergem, O. K., & Roe, A. (2008). *PISA+: Lærings-og undervisningsstrategier i skolen*. Oslo: Norges Forskningsråd.

Kline, R. B. (2011). *Principles and practice of Structural Equation Modeling* (3rd ed.). New York: The Guildford Press.

Kolenikov, S., & Bollen, K. A. (2012). *Testing Negative Error Variances: Is a Heywood Case a Symptom of Misspecification? Sociological Methods & Research* (Vol. 41). http://doi.org/10.1177/0049124112442138

Lacker, J. M., Carosella, A. M., & Feuerstein, M. (1996). Pain expectancies, pain, and functional self-efficacy expectancies as determinants of disability in patients with chronic low back disorders. *Journal of Consulting and Clinical Psychology*, *64*(1), 212–220. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8907101

Langfeldt, G. (2015). Skolens kvalitet skapes lokalt. In *Skolens kvalitet skapes lokalt. Presentasjon av funn fra forskningsprosjektet "Lærende Regioner."* Oslo: Fagbokforlaget.

Lee, C., & Bobko, P. (1994). Self-efficacy beliefs: Comparison of five measures. *Journal of Applied Psychology*, *79*(3), 364–369. http://doi.org/10.1037/0021-9010.79.3.364

Lei, P.-W., & Wu, Q. (2007). Introduction to Structural Equation Modeling: Issues and Practical Considerations. *Educational Measurement: Issues and Practices (ITEMS Module)*, *26*(3), 33–43. http://doi.org/10.1111/j.1745-3992.2007.00099.x

Lei, P.-W., & Wu, Q. (2012). Estimation in Structural Equation Modeling. In R. G. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (Paperback, pp. 164–180). New

York: The Guildford Press.
Lent, R. W., Lopez, F. G., & Bieschke, K. J. (1993). Predicting mathematics-related choice and success behaviors: Test of an expanded social cognitive model. *Journal of Vocational Behavior*, *42*, 223–236.
Lent, R. W., Lopez, F. G., Brown, S. D., & Gore Jr., P. A. (1996). Latent Structure of the Sources of Mathematics Self-Efficacy. *Journal of Vocational Behavior*, *49*, 292–308.
Lewis, J. L., Ream, R. K., Bocian, K. M., Cardullo, R. A., Hammond, K. A., & Fast, L. A. (2012). Con cariño: Teacher caring, math self-efficacy, and math achievement among hispanic english learners. *Teachers College Record*, *114*(7), 1–42.
Lindberg, S. M., Shibley Hyde, J., Peterson, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin*, *136*(6), 1123–1135.
Lindsley, D. H., Brass, D. J., & Thomas, J. B. (1995). Efficacy-Performing Spirals: a Multilevel Perspective. *Academy of Management Review*, *20*(3), 645–678. http://doi.org/10.5465/AMR.1995.9508080333
Little, T. D. (1997). Mean and Covariance Structures (MACS) Analyses of Cross-Cultural Data: Practical and Theoretical Issues. *Multivariate Behavioral Research*, *32*(1), 53–76. http://doi.org/10.1207/s15327906mbr3201
Little, T. D. (2013). *Longitudinal Structural Equation Modeling*. New York: The Guildford Press.
Little, T. D., Cunningham, W. a., Shahar, G., & Widaman, K. F. (2002). To Parcel or Not to Parcel: Exploring the Question, Weighing the Merits. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 151–173. http://doi.org/10.1207/S15328007SEM0902_1
Lloyd, J. E. V., Walsh, J., & Yailagh, M. S. (2005). Sex Differences in Performance Attributions, Self-Efficacy, and Achievement in Mathematics: If I' m So Smart, Why Don't I Know It? *Canadian Journal of Education*, *28*(3), 384–408.
Locke, E. A., Frederick, E., Lee, C., & Bobko, P. (1984). Effect of self-efficacy, goals, and task strategies on task performance. *Journal of Applied Psychology*, *69*(2), 241–251. http://doi.org/10.1037/0021-9010.69.2.241
Lodewyk, K. R., & Winne, P. H. (2005). Relations Among the Structure of Learning Tasks, Achievement, and Changes in Self-Efficacy in Secondary Students. *Journal of Educational Psychology*, *97*(1), 3–12. http://doi.org/10.1037/0022-0663.97.1.3
Malmberg, L.-E., & Little, T. D. (2007). Profiles of ability, effort, and difficulty: Relationships with worldviews, motivation and adjustment. *Learning and Instruction*, *17*(6), 739–754. http://doi.org/10.1016/j.learninstruc.2007.09.014
Malmberg, L.-E., Pakarinen, E., Vasalampi, K., & Nurmi, J.-E. (2015). Students` school performance, task-focus and situation-specific motivation. *Learning and Instruction*.
Marsh, H. W., Byrne, B. M., & Shavelson, R. J. (1988). A multifaceted academic self-concept: Its hierarchical structure and its relation to academic achievement. *Journal of Educational Psychology*, *80*(3), 366–380. http://doi.org/10.1037//0022-0663.80.3.366
Marsh, H. W., Dowson, M., Pietsch, J., & Walker, R. (2004). Why Multicollinearity Matters: A Reexamination of Relations Between Self-Efficacy, Self-Concept, and Achievement. *Journal of Educational Psychology*, *96*(3), 518–522. http://doi.org/10.1037/0022-0663.96.3.518
Marsh, H. W., Hau, K.-T., Balla, J. R., & Grayson, D. (1998). Is More Ever Too Much ? The Number of Indicators Per Factor in Confirmatory Factor Analysis. *Multivariate Behavioral Research*, *33*(2), 181–220. http://doi.org/10.1207/s15327906mbr3302

Marsh, H. W., Scalas, L. F., & Nagengast, B. (2010). Longitudinal tests of competing factor structures for the Rosenberg Self-Esteem Scale: Traits, ephemeral artifacts, and stable response styles. *Psychological Assessment*, *22*(2), 366–381. http://doi.org/10.1037/a0019225

Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2013). Academic self-concept, interest, grades, and standardized test scores: reciprocal effects models of causal ordering. *Child Development*, *76*(2), 397–416. http://doi.org/10.1111/j.1467-8624.2005.00853.x

Martin, A. J. (2003). The student motivation scale: Further testing of an instrument that measures school students` motivation. *Australian Journal of Education*, *47*(I), 88–106.

Martin, A. J., Anderson, J., Bobis, J., Way, J., & Vellar, R. (2012). Switching on and switching off in mathematics: An ecological study of future intent and disengagement among middle school students. *Journal of Educational Psychology*, *104*(1), 1–18. http://doi.org/10.1037/a0025988

Matsui, T., Matsui, K., & Ohnishi, R. (1990). Mechanisms Underlying Math Self-Efficacy Learning of College Students. *Journal of Vocational Behavior*, *37*(2), 225–238.

McConney, A., & Perry, L. B. (2010). Socioeconomic status, self-efficacy, and mathematics achievement in Australia: a secondary analysis. *Educational Research for Policy and Practice*, *9*(2), 77–91. http://doi.org/10.1007/s10671-010-9083-4

McConville, S. A., & Lane, A. M. (2006). Using on-line video clips to enhance self-efficacy toward dealing with difficult situations among nursing students. *Nurse Education Today*, *26*(3), 200–208. http://doi.org/10.1016/j.nedt.2005.09.024

Meece, J. L., Wigfield, A., & Eccles, J. S. (1990). Predictors of math anxiety and its influence on young adolescents' course enrollment intentions and performance in mathematics. *Journal of Educational Psychology*, *82*(1), 60–70. http://doi.org/10.1037/0022-0663.82.1.60

Mega, C., Ronconi, L., & De Beni, R. (2014). What makes a good student? How emotions, self-regulated learning, and motivation contribute to academic achievement. *Journal of Educational Psychology*, *106*(1), 121–131. http://doi.org/10.1037/a0033546

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods*, *10*(3), 259–284. http://doi.org/10.1037/1082-989x.10.3.259

Midgley, C., Feldlaufer, H., & Eccles, J. S. (1989). Change in teacher efficacy and student self- and task-related beliefs in mathematics during the transition to junior high school. *Journal of Educational Psychology*, *81*(2), 247–258. http://doi.org/10.1037//0022-0663.81.2.247

Midgley, C., Kaplan, A., Middleton, M., Maehr, M. L., Urdan, T. C., Anderman, L. H., … Roeser, R. (1998). The development and validation of scales assessing students' achievement goal orientations. *Contemporary Educational Psychology*, *23*(2), 113–131. http://doi.org/10.1006/ceps.1998.0965

Midgley, C., Maehr, M. L., Hruda, L. Z., Anderman, E. M., Anderman, L. H., Freeman, K. E., … Urdan, T. C. (2000). *Manual for the Patterns of Adaptive Learning Scales*. University of Michigan.

Miliaressis, E., & Rompre, P. P. (1987). Effects of concomitant motor reactions on the measurement of rewarding efficacy of brain stimulation. *Behavioral Neuroscience*, *101*(6), 827–831. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/3426797

Morin, A. J. S., Marsh, H. W., Nagengast, B., & Scalas, L. F. (2014). Doubly Latent Multilevel

Analyses of Classroom Climate: An Illustration. *The Journal of Experimental Education*, *82*(2), 143–167. http://doi.org/10.1080/00220973.2013.769412
Muis, K. R. (2004). Personal Epistemology and Mathematics: A Critical Review and Synthesis of Research. *Review of Educational Research*, *74*(3), 317–377.
Muis, K. R., & Duffy, M. C. (2013). Epistemic climate and epistemic change: Instruction designed to change students' beliefs and learning strategies and improve achievement. *Journal of Educational Psychology*, *105*(1), 213–225. http://doi.org/10.1037/a0029690
Multon, K. D., Brown, S. D., & Lent, R. W. (1991). Relation of self-efficacy beliefs to academic outcomes: A meta-analytic investigation. *Journal of Counseling Psychology*, *38*(1), 30–38. http://doi.org/10.1037//0022-0167.38.1.30
Muthén, L. K., & Muthén, B. O. (2012). *Mplus User`s Guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.
Nicholls, J. G. (1978). The Development of the Concepts of Effort and Ability, Perception of Academic Attainment and the Understanding That Difficult Tasks Require More Ability. *Child Development*, *49*(3), 800–814.
Norwegian Directorate for Education. (2016). Nasjonale prøver. Retrieved April 28, 2016, from http://www.udir.no/Vurdering/Nasjonale-prover/
Norwich, B. (1987). Self-efficacy and mathematics achievement: A study of their relation. *Journal of Educational Psychology*, *79*(4), 384–387.
Opplæringslova (1998). Lov om grunnskolen og den vidaregående opplæringa (LOV-1998-07-17-61). Retrieved from https://lovdata.no/lov/1998-07-17-61
Pajares, F. (1992). Teachers ' Beliefs and Educational Research : Cleaning Up a Messy Construct. *Review of Educational Research*, *62*(3), 307–332.
Pajares, F. (1996a). Self-Efficacy Beliefs and Mathematical Problem-Solving of Gifted Students. *Contemporary Educational Psychology*, *21*(4), 325–44. http://doi.org/S0361476X96900259 [pii]
Pajares, F. (1996b). Self-efficacy beliefs in academic settings. *Review of Educational Research*, *66*(4), 543–578.
Pajares, F. (2002). Gender and Perceived Self-Efficacy in Self-Regulated Learning. *Theory into Practice*, *41*(2), 116–125. http://doi.org/10.1207/s15430421tip4102
Pajares, F., & Graham, L. (1999). Self-Efficacy, Motivation Constructs, and Mathematics Performance of Entering Middle School Students. *Contemporary Educational Psychology*, *24*(2), 124–139. http://doi.org/10.1006/ceps.1998.0991
Pajares, F., & Kranzler, J. (1995). Self-Efficacy Beliefs and General Mental Ability in Mathematical Problem-Solving. *Contemporary Educational Psychology*, 426–443.
Pajares, F., & Miller, M. D. (1994). Role of self-efficacy and self-concept beliefs in mathematical problem solving: A path analysis. *Journal of Educational Psychology*, *86*(2), 193–203.
Pajares, F., & Miller, M. D. (1995). Mathematics self-efficacy and mathematics performances: The need for specificity of assessment. *Journal of Counseling Psychology*, *42*(2), 190–198.
Pajares, F., & Valiante, G. (2001). Gender Differences in Writing Motivation and Achievement of Middle School Students: A Function of Gender Orientation? *Contemporary Educational Psychology*, *26*(3), 366–381. http://doi.org/10.1006/ceps.2000.1069
Pampaka, M., Kleanthous, I., Hutcheson, G. D., & Wake, G. (2011). Measuring mathematics self-efficacy as a learning outcome. *Research in Mathematics Education*, *13*(2), 169–190. http://doi.org/10.1080/14794802.2011.585828

Penner, A. M. (2008). Gender Differences in Extreme Mathematical Achievement : An International Perspective on Biological and Social Factors. *American Journal of Sociology*, *114*.
Peters, M. L. (2013). Examining the relationships among classroom climate, self-efficacy, and achievement in undergraduate mathematics: a multi-level analysis. *International Journal of Science and Mathematics Education*, *11*, 459–480.
Phan, H. P. (2012). The Development of English and Mathematics Self-Efficacy: A Latent Growth Curve Analysis. *The Journal of Educational Research*, *105*(3), 196–209. http://doi.org/10.1080/00220671.2011.552132
Phan, H. P. (2013). Self-Efficacy, Reflection, and Achievement: A Short-Term Longitudinal Examination. *The Journal of Educational Research*, *0*, 1–13.
Pietsch, J., Walker, R., & Chapman, E. (2003). The relationship among self-concept, self-efficacy, and performance in mathematics during secondary school. *Journal of Educational Psychology*, *95*(3), 589–603. http://doi.org/10.1037/0022-0663.95.3.589
Pintrich, P. R., & de Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology*, *82*(1), 33–40. http://doi.org/10.1037//0022-0663.82.1.33
Pintrich, P. R., Smith, D. A. F., Garcia, T., & McKeachie, W. J. (1993). Reliability and predictive validity of the motivated strategies for learning questionnaire (MSLQ). *Educational and Psychological Measurement*, *53*, 801–813. http://doi.org/0803973233
Randhawa, B. S., Beamer, J. E., & Lundberg, I. (1993). Role of mathematics self-efficacy in the structural model of mathematics achievement. *Journal of Educational Psychology*, *85*(1), 41–48.
Ray, R., Gornick, J. C., & Schmitt, J. (2009). *Parental leave policies in 21 countries: Assessing generosity and gender equality*. Washington.
Rhodewalt, F. (1994). Conceptions of Ability, Achievement Goals, and Individual Differences in Self-Handicapping Behavior: On the Application of Implicit Theories. *Journal of Personality*, *62*(March 1994), 67–76.
Rimm-Kaufman, S. E., Baroody, A. E., Larsen, R. A. A., Curby, T. W., & Abry, T. (2015). To What Extent Do Teacher – Student Interaction Quality and Student Gender Contribute to Fifth Graders ' Engagement in Mathematics Learning? *Journal of Educational Psychology*, *107*(1), 170–185.
Rogers, L. Q., Courneya, K. S., Verhulst, S., Markwell, S., Lanzotti, V., & Shah, P. (2006). Exercise barrier and task self-efficacy in breast cancer patients during treatment. *Supportive Care in Cancer : Official Journal of the Multinational Association of Supportive Care in Cancer*, *14*(1), 84–90. http://doi.org/10.1007/s00520-005-0851-2
Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the Fit of Structural Equation Models : Tests of Significance and Descriptive Goodness-of-Fit Measures. *Methods of Psychological Research Online*, *8*(2), 23–74.
Schmitz, B. (2006). Advantages of studying processes in educational research. *Learning and Instruction*, *16*(5), 433–449. http://doi.org/10.1016/j.learninstruc.2006.09.004
Schmitz, B., & Wiese, B. S. (2006). New perspectives for the evaluation of training sessions in self-regulated learning: Time-series analyses of diary data. *Contemporary Educational Psychology*, *31*(1), 64–96. http://doi.org/10.1016/j.cedpsych.2005.02.002
Scholz, U., Dona, B. G., Sud, S., & Schwarzer, R. (2002). Is general self-efficacy a universal

construct? Psychometric findings from 25 countries. *European Journal of Psychological Assessment*, *18*(3), 242–251. Retrieved from http://search.proquest.com/docview/614367582?accountid=38978
Schraw, G., Potenza, M. T., & Nebelsick-Gullet, L. (1993). Constraints on the calibration of performance. *Contemporary Educational Psychology*, *18*(4), 455–463.
Schukajlow, S., Leiss, D., Pekrun, R., Blum, W., Müller, M., & Messner, R. (2012). Teaching methods for modelling problems and students' task-specific enjoyment, value, interest and self-efficacy. *Educational Studies in Mathematics*, *79*(2), 215–237.
Schulz, W. (2005). Mathematics Self-Efficacy and Student Expectations. Results from PISA 2003. In *Annual meetings of the American Educational Research Association* (pp. 11–15). Montreal.
Schunk, D. H. (1981). Modeling and attributional effects on children's achievement: A self-efficacy analysis. *Journal of Educational Psychology*, *73*(1), 93–105.
Schunk, D. H. (1989). Self-efficacy and achievement behaviors. *Educational Psychology Review*, *1*(3), 173–208. http://doi.org/10.1080/00461528409529281
Schunk, D. H. (1991). Self-efficacy and academic motivation. *Educational Psychologist*, *26*, 207–231.
Schunk, D. H., & Lilly, M. W. (1984). Sex Differences in Self-Efficacy and Attributions: Influence of Performance Feedback. *The Journal of Early Adolescence*, *4*(3), 203–213. http://doi.org/10.1177/0272431684043004
Schunk, D. H., & Miller, S. D. (2002). Self-efficacy and adolescents` motivation. In F. Pajares & T. Urdan (Eds.), *Academic motivation of adolescents* (pp. 29–52). Greenwich, Connecticut: Information Age Publishing Inc. Retrieved from https://www.uky.edu/~eushe2/Pajares/AdoEd2.html
Schunk, D. H., & Pajares, F. (2010). Self-Efficacy Beliefs. In P. Peterson, E. L. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education* (3rd ed., pp. 668–672). New York: Elsevier Ltd.
Schunk, D. H., & Zimmerman, B. J. (2007). Influencing Children's Self-Efficacy and Self-Regulation of Reading and Writing Through Modeling. *Reading & Writing Quarterly*, *23*(1), 7–25. http://doi.org/10.1080/10573560600837578
Schweinle, A., Berg, P. J., & Sorenson, A. R. (2013). Preadolescent perceptions of challenging and difficult course activities and their motivational distinctions. *Educational Psychology*, *33*(7), 797–816. http://doi.org/10.1080/01443410.2013.785049
Schweinle, A., & Mims, G. A. (2009). Mathematics self-efficacy: stereotype threat versus resilience. *Social Psychology of Education*, *12*(4), 501–514. http://doi.org/10.1007/s11218-009-9094-2
Schweinle, A., Turner, J. C., & Meyer, D. K. (2006). Striking the right balance: Students' motivation and affect in elementary mathematics. *The Journal of Educational Research*, *99*(5).
Shea, C. M., & Howell, J. M. (2000). Efficacy-Performance Spirals : An Empirical Test. *Journal of Management*, *26*(4), 791–812.
Shih, S.-S., & Alexander, J. M. (2000). Interacting Effects of Goal Setting and Self- or Other-Referenced Feedback on Children' s Development of Self-Efficacy and Cognitive Skill Within the Taiwanese Classroom. *Journal of Educational Psychology*, *92*(3), 536–543. http://doi.org/10.1037//0022-0663.92
Skaalvik, E. M., & Skaalvik, S. (2009). Elevens opplevelse av skolen: Sentrale sammenhenger og utvikling med alder. *Spesialpedagogikk*, *8*, 36–47.
Skaalvik, E. M., & Valås, H. (1999). Relations Among Achievement, Self-Concept, and

Motivation in Mathematics and Language Arts: A Longitudinal Study. *The Journal of Experimental Education*, *67*(2), 135–149.
Snyder, K. E., Malin, J. L., Dent, A. L., & Linnenbrink-Garcia, L. (2014). The message matters: The role of implicit beliefs about giftedness and failure experiences in academic self-handicapping. *Journal of Educational Psychology*, *106*(1), 230–241.
Sørlie, K., & Söderlund, G. (2015). Regionale forskjeller i mestringsforventninger. In G. Langfeldt (Ed.), *Skolens kvalitet skapes lokalt. Presentasjon av funn fra forskningsprosjektet "Lærende Regioner."* Oslo: Fagbokforlaget.
Stajkovic, A. D., & Luthans, F. (1998). Self-efficacy and work-related performance: A meta-analysis. *Psychological Bulletin*, *124*(2), 240–261. http://doi.org/10.1037//0033-2909.124.2.240
Stajkovic, A. D., & Sommer, S. M. (2000). Self-Efficacy and Causal Attributions : Direct and Reciprocal Links. *Journal of Applied Social Psychology*, *30*(4), 707–737.
Statistics Norway. (2017). *Elevar i grunnskolen, 14.december 2017*. Retrieved from https://www.ssb.no/utdanning/statistikker/utgrs
Stipek, D. J. (1981). Children's perceptions of their own and their classmates' ability. *Journal of Educational Psychology*, *73*(3), 404–410. http://doi.org/10.1037/0022-0663.73.3.404
Street, K. E. S., Malmberg, L.-E., & Stylianides, G. J. (2017a). Level, strength, and facet-specific self-efficacy in mathematics test performance. *ZDM - Mathematics Education*, *49*(3), 379–395. http://doi.org/10.1007/s11858-017-0833-0
Street, K. E. S., Malmberg, L.-E., & Stylianides, G. J. (2017b). Self-efficacy and Mathematics Performance: Reciprocal Relationships. In T. Dooley & G. Gueudet (Eds.), *Proceedings of the Tenth Congress of the European Society for Research in Mathematics Education* (pp. 1186–1193).
Stylianides, A. J., & Stylianides, G. J. (2014). Impacting positively on students' mathematical problem solving beliefs: An instructional intervention of short duration. *The Journal of Mathematical Behavior*, *33*, 8–29. http://doi.org/10.1016/j.jmathb.2013.08.005
Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). London: Pearson.
Talsma, K., Schüz, B., Schwarzer, R., & Norris, K. (2018). I believe, therefore I achieve (and vice versa): A meta-analytic cross-lagged panel analysis of self-efficacy and academic performance. *Learning and Individual Differences*, *61*(January), 136–150. http://doi.org/10.1016/j.lindif.2017.11.015
Throndsen, I. (2011). Self-regulated learning of basic arithmetic skills: A longitudinal study. *British Journal of Educational Psychology*, *81*(4), 558–578.
Usher, E. L. (2009). Sources of Middle School Students ' Self-Efficacy in Mathematics : A Qualitative Investigation. *American Educational Research Journal*, *46*(1), 275–314.
Usher, E. L., & Pajares, F. (2006). Sources of academic and self-regulatory efficacy beliefs of entering middle school students. *Contemporary Educational Psychology*, *31*(2), 125–141. http://doi.org/10.1016/j.cedpsych.2005.03.002
Usher, E. L., & Pajares, F. (2008). Sources of Self-Efficacy in School: Critical Review of the Literature and Future Directions. *Review of Educational Research*, *78*(4), 751–796.
Usher, E. L., & Pajares, F. (2009). Sources of self-efficacy in mathematics: A validation study. *Contemporary Educational Psychology*, *34*(1), 89–101.
Valentine, J. C., & DuBois, D. L. (2005). Effects of self-beliefs on academic achievement and vice versa. In H. Marsh, R. Craven, & D. McInerney (Eds.), *New Frontiers for self research* (pp. 53–77). Information Age Publishing Inc.

Valentine, J. C., Dubois, D. L., & Cooper, H. M. (2004). The Relation Between Self-Beliefs and Academic Achievement: A Meta-Analytic Review. *Educational Psychologist*, *39*(2), 111–133. http://doi.org/10.1207/s15326985ep3902

Vermunt, J. D., & Verloop, N. (1999). Congruence and friction between learning and teaching. *Learning and Instruction*, *9*(3), 257–280.

Wigfield, A., Eccles, J. S., & Rodriguez, D. (1998). The Development of Children' s Motivation in School Contexts. *Review of Research in Education*, *23*, 73–118.

Wigfield, A., Eccles, J. S., Yoon, K. S., Harold, R. D., Arbreton, A. J. A., Freedman-Doan, C., & Blumenfeld, P. C. (1997). Change in children's competence beliefs and subjective task values across the elementary school years: A 3-year study. *Journal of Educational Psychology*, *89*(3), 451–469. http://doi.org/10.1037/0022-0663.89.3.451

Wigfield, A., & Tonks, S. (2002). Adolescents` expectancies for success and achievement task values during the middle and high school years. In F. Pajares & T. Urdan (Eds.), *Academic motivation of adolescents* (pp. 53–82). Greenwich, Connecticut: Information Age Publishing Inc.

Williams, T., & Williams, K. (2010). Self-efficacy and performance in mathematics: Reciprocal determinism in 33 nations. *Journal of Educational Psychology*, *102*(2), 453–466. http://doi.org/10.1037/a0017271

Winfrey, M. L., & Weeks, D. L. (1993). Effects of self-modeling on self-efficacy and balance beam performance. *Percept Mot Skills*, *77*(3 Pt 1), 907–913. Retrieved from

Yailagh, M. S., Lloyd, J. E. V., & Walsh, J. (2009). The causal relationship between attribution styles, mathematics self-efficacy beliefs, gender differences, goal setting and math achievement of school children. *Journal of Education and Psychology*, *3*(2), 95–114.

Yetkin Özdemir, İ. E., & Pape, S. J. (2013). The Role of Interactions Between Student and Classroom Context in Developing Adaptive Self-Efficacy in One Sixth-Grade Mathematics Classroom. *School Science and Mathematics*, *113*(5), 248–258.

Zeldin, A. L., & Pajares, F. (2000). Against the Odds: Self-Efficacy Beliefs of Women in Mathematical, Scientific, and Technological Careers. *American Educational Research Journal*, *37*(1), 215–246.

Zimmerman, B. J. (1990). Self-Regulated Learning and Academic Achievement : An Overview. *Educational Psychologist*, *25*(1), 3–17.

Zimmerman, B. J. (2000). Self-Efficacy: An Essential Motive to Learn. *Contemporary Educational Psychology*, *25*(1), 82–91. http://doi.org/10.1006/ceps.1999.1016

Zimmerman, B. J., Bandura, A., & Martinez-Pons, M. (1992). Self-Motivation for Academic Attainment: The Role of Self-Efficacy Beliefs and Personal Goal Setting, *29*(3), 663–676.

Zimmerman, B. J., & Kitsantas, A. (2005). Homework practices and academic achievement: The mediating role of self-efficacy and perceived responsibility beliefs. *Contemporary Educational Psychology*, *30*(4), 397–417. http://doi.org/10.1016/j.cedpsych.2005.05.003

Zimmerman, B. J., & Martinez-Pons, M. (1990). Student differences in self-regulated learning: Relating grade, sex, and giftedness to self-efficacy and strategy use. *Journal of Educational Psychology*, *82*(1), 51–59. http://doi.org/10.1037/0022-0663.82.1.51

Zimmerman, B. J., & Ringle, J. (1981). Effects of model persistence and statements of confidence on children's self-efficacy and problem solving. *Journal of Educational Psychology*, *73*(4), 485–493. http://doi.org/10.1037//0022-0663.73.4.485

Appendix A: Submitted article

Level, strength, and facet-specific self-efficacy in mathematics test performance

Karin Elisabeth Sørlie Street [a b], Lars-Erik Malmberg [a] and Gabriel J. Stylianides [a]

[a] Department of Education, University of Oxford

[b] Department of Teacher Education and Sport, University College of Sogn and Fjordane

Abstract:
Students' self-efficacy expectations (SEE) in mathematics are associated with their engagement and learning experiences. Going beyond previous operationalisations of SEE we propose a new instrument that takes into account not only *facet-specificity* (expectations related to particular competences or skills) and *strength* (confidence of the expectations), but also *level* (perceived task difficulty) of these expectations as proposed by Bandura (1997, 2006). In particular, we included level-specific items referring to perceived difficulty on a subsequent national test in mathematics. In total 756 Norwegian grade 5, 8, and 9 students completed the "Self-Efficacy Gradations of Difficulty Questionnaire." We fitted plausible multitrait-multimethod models using structural equation models. The best fitting model included three factors representing levels of perceived difficulty, and a-priori specified correlated uniquenesses representing four facets. The facets related to problem solving or students' self-regulation skills during the test in order to accomplish the following: (1) complete a certain number of problems, (2) solve tasks of a certain challenge, (3) concentrate, and (4) not give up for a certain amount of time. The results indicated that three correlated constructs representing levels of SEE are associated with scores on national tests in mathematics, and that the strongest association is between national test scores and medium level SEE. Taking level (difficulty) into account broadens our understanding of the self-efficacy construct, and allows investigation into differential relationships between SEE and performance.

Keywords: self-efficacy, levels of difficulty, expectations, mathematics, national tests

Level, strength, and facet-specific self-efficacy in mathematics test performance

1. Introduction

Self-efficacy expectations (SEE) are related to key learning behaviours and outcomes such as individuals' motivation and engagement (Martin et al. 2012), goal setting behaviours (Bong 2009), persistence (Multon et al. 1991), and physiological experiences (Meece et al. 1990). In mathematics in particular, SEE have been linked with perseverance (Zeldin and Pajares 2000), performance outcomes (Schulz 2005), attitudes towards mathematics and choice of maths-related courses and careers (Hackett and Betz 1989). Previous studies have used a wide range of instruments to measure SEE. Existing instruments have incorporated *facet-specificity* (expectations related to particular competences or skills) and *strength* (confidence of the expectations), while *levels* (perceived task difficulty) of mathematics SEE have seemingly not been included.

In this paper, we contribute to existing research by proposing a new instrument of SEE that takes into account not only strength and facet-specificity, but also level of SEE as proposed by Bandura (1997, 2006). Taking level into account broadens our understanding of the self-efficacy construct, because it allows the distinction between SEE for different levels of perceived difficulty, as well as investigation into the relationship between SEE and performance outcomes for each level. Specifically, we tested the structural validity of our proposed multidimensional instrument, and we investigated how level and facet-specificity of SEE are related to subsequent performance on national tests in mathematics. To this end, we used a sample of 756 Norwegian grade 5, 8, and 9 students who completed what we call the "Self-Efficacy Gradations of Difficulty Questionnaire."

In the following sections, we first review Bandura's proposals about the meaning of SEE. We then outline facet-specificity, strength, and level of SEE and discuss how these have been operationalised and measured in previous studies. After that we present our model that integrates facet-specificity, strength, and level of mathematics SEE.

2. Self-efficacy

The construct of self-efficacy was introduced by Albert Bandura, who defined s*elf-efficacy expectations* as "beliefs in one's capabilities to organize and execute the courses of action required to produce given attainments" (1997, p. 3). Researchers of SEE such as Schunk and Pajares (2010), Zimmerman (2000), and Bong and Skaalvik (2003) all refer to versions of the original conceptualisation proposed by Bandura. Self-efficacy is either referred to as a belief (Multon et al. 1991; Schunk and Pajares 2010) or an expectation

(Hackett and Betz 1989; Schukajlow et al. 2012). We use the term *expectation* to emphasise the focus on prospective future attainments.

SEE differ from other self-beliefs in their relation to specific, future challenges or tasks and their relative context specificity (Bong and Skaalvik 2003; Pajares 1996b; Valentine et al. 2004). Researchers further emphasise a mastery criterion of performance in relation to SEE, which stands in contrast to the normative comparisons proposed as an important source of self-concept (Bong and Skaalvik 2003; Zimmerman 2000). SEE also differ from outcome expectancies: while SEE are related to individuals' perceived ability to perform activities, outcome expectancies are related to anticipated consequences of such performances (Zimmerman 2000).

SEE vary according to three dimensions: generality, strength, and level (Bandura 1997; Pajares 1996b; Zimmerman 2000). *Generality* refers to the degree of specificity. For instance, SEE for mathematics are more general as compared to SEE for specific facets of doing mathematics, such as solving problems. In this paper, we use the term *facet-specificity* in place of "generality" given that our theoretical model includes specific facets of taking a mathematics test. *Strength* refers to the amount of certainty about performance on a task (such as doing well on a national test in mathematics) or carrying out specific facets of test taking (such as persevering). *Level* refers to the degree of performance demand, difficulty, or challenge included in the task. For instance, level can refer to SEE for tasks of increasing difficulty or conducted under increasingly difficult circumstances. Next we elaborate on each of these three dimensions.

2.1. Facet-specificity

SEE are relatively specific constructs (Bandura 1997; Bong 2001; Schukajlow et al. 2012), where individuals form their judgment based on an assessment of prospective future tasks and situations. According to this conceptualisation, global constructs such as general self-efficacy--referring to individuals' overall confidence across a range of domains, situations, and tasks (e.g., Scholz et al. 2002)--are unlikely to address participants' subject-specific SEE (Pajares and Miller 1995). An example item for general self-efficacy is "*Thanks to my resourcefulness, I can handle unforeseen situations"*. Bong (2001) argued SEE are subject-specific, yet moderately correlated across subjects. Bong found that older students' SEE were more subject-specific than those of younger students, demonstrated in lower cross-subject correlations of SEE (median value 0.55 for middle school sample, 0.42 for high school sample). SEE might also vary across different facets within an

activity domain (Bandura 2006), and Pajares (1996b) argued that even subject-specific measures can fail to prompt participants to have a specific task or activity in mind when responding.

Bandura (2006, pp. 324, 326) proposed instruments to measure facet-specificity of SEE, such as problem-solving self-efficacy and self-efficacy for self-regulated learning. To gauge participants' problem solving-specific SEE, they were asked to rate how certain they were that they could solve a certain proportion of academic problems, e.g., 10% of the problems, 20% of the problems, and so on up to 100% of the problems. Participants' self-efficacy for self-regulation was gauged by asking how certain they were that they could concentrate on school subjects during class, and persevere in their school-work.

Within the domain of mathematics, SEE may be specified in regard to learning mathematics, sub-areas of mathematics, mathematical competencies, or specific mathematical tasks (Schukajlow et al. 2012). Hackett and Betz (1989) operationalised mathematics SEE in a wide sense, including confidence to solve mathematical problems, performance of everyday mathematical tasks, and doing well on mathematics-related courses.

Pajares (1996a) found that task-specific mathematics SEE predicted performance on those particular tasks for gifted ($\beta = .46$) and regular education ($\beta = .39$) students. For both samples, SEE for self-regulated learning predicted mathematics SEE ($\beta = .42$ and $\beta = .37$, respectively), but not performance. Mathematics SEE were measured by first displaying the tasks to the participants, before asking them to rate their confidence in solving them. This strategy enables high correspondence between SEE and the performance measures, and has been commonly applied (e.g., Dowling 1978; Schukajlow et al. 2012). SEE for self-regulated learning were measured with items from the Children's Self-Efficacy for Self-Regulated Learning scale (Bandura 2006), which is related to general schoolwork (e.g., organise and plan schoolwork, and concentrate during class). In the study by Pajares (1996a) the measure for problem solving SEE was more specific than the measure for SEE for self-regulation, while in our study items related to problem solving SEE and SEE for self-regulation are equally specific.

Zimmerman and Kitsansas (2005) included test taking as one facet of their Self-Efficacy for Learning Form, designed to assess students' confidence to cope with a range of academic learning or performance challenges. The authors included perceived skills in self-regulation to measure SEE for taking tests (example item: *When you feel anxious*

during an exam and have trouble controlling information, can you relax and concentrate well enough to remember it?).

Consistently with the literature, we propose an instrument that focuses on a prospective test-taking situation. We regard problem-solving (complete a certain number of problems; solve tasks of a certain challenge) as well as self-regulation (concentrate; not give up) as relevant facets of the test-taking situation. By including facets related to the test-taking situation, we take into account also empirical research which has demonstrated that the relationship between self-beliefs and outcome measures is generally stronger with increasing specificity and correspondence between the measures (Valentine et al. 2004). We use items related to students' perceived capability of performing on the test, which differs from the outcomes they expect from their performance. SEE for each facet includes the "strength" of SEE for each item related to the specific facet. We elaborate on the notion of strength next.

2.2. Strength

Strength of SEE is related to the amount of individuals' confidence in their own capabilities. Bandura (1997) contends that strong SEE are more resilient to change than are weakly held expectations. A majority of investigations into the relationship between SEE and performance outcomes include strength of SEE as a predictor of the dependent variable. Strength of SEE generally predicts performance, where individuals reporting stronger SEE (more confident) tend to perform better on subsequent tests in related activities. Multon et al. (1991) found an average effect size of 0.38 across 36 studies, while Pajares (1996b) reported that direct effects ranged from 0.35 to 0.55 in path analytic studies.

According to Lee and Bobko (1994), strength of SEE has commonly been estimated by combining the confidence ratings from all the items included in the measure. In their study they compared different ways of measuring SEE, including pure strength and pure level measures. They asked participants to indicate whether they could perform a task at a certain level (yes/no), and to indicate the strength of their confidence on a scale from 0-10. The authors found that combining the strength scores from items that were answered "yes" led to higher convergent and predictive validities than pure strength or pure level measures yielded.

In our study we seek to disentangle the effects of facet-specificity and levels of difficulty. Similarly to Lee and Bobko (1994), we propose a composite measure. We go

beyond their study, though, in that we specify structural equation models in which strength information is included in relation to latent level constructs and latent facet constructs. Strength of SEE is related to each individual item in the measure, which are also related to a specific facet and level of difficulty. Strength of SEE is thus implicitly included in SEE for each of the four facets we identified and each of the levels of difficulty (easy, medium, or hard). Furthermore, overall strength of SEE can be estimated through combining the confidence ratings for all items.

2.3. Level

Bandura (1997) emphasises that SEE are not the same as people`s beliefs about the skills they possess, but a measure of what they believe they can do with those skills. Accordingly, situational conditions are important, as they represent different levels of demand or impediments to success in the particular domain or task. Bandura argues that instruments referring to a single level of task demand might fail to distinguish between individuals who differ in their SEE for tasks of different levels of difficulty, and notes that "challenges may be graded in terms of level of ingenuity, exertion, accuracy, productivity, threat, or self-regulation required, just to mention a few dimensions of performance demands" (1997, p. 43). We found only a few studies that included SEE level in their measure or analyses. Furthermore, practices for analyses of level of SEE are not consistent.

One example is the Memory Self-Efficacy Questionnaire, a multidimensional self-efficacy scale devised by Berry et al. (1989). According to Bandura, this measure responds well to the theoretical conceptions and guidelines from self-efficacy theory and methodology. It includes items related to 10 different types of memory task, and five different levels of demand for each task. An example item is as follows: *If I heard it twice, I could remember 12 items from a friend's grocery list of 12 items, without taking any list with me to the store*. The different levels of demand for this task include progressively easier gradations of difficulty, in regard to how many grocery items the participant can remember. For each of the different types of memory task, participants are asked to circle either "No" or "Yes". If they respond "Yes" they then rate their confidence on a scale from 10% to 100%, in 10-unit increments. For each of the 10 memory tasks the authors analysed participants' sum of "Yes" responses as self-efficacy *level,* and their average confidence across five levels as self-efficacy *strength*. The authors thus analysed

SEE for each memory task separately, and used intercorrelation estimates to assess task specificity post hoc.

The above measure shares similarities with the one proposed in our study in that it includes several types of tasks for each facet and of different levels of difficulty. At the same time, the two studies differ in conceptualisation and analyses. Berry et al. (1989) combined "Yes" responses across levels of difficulty to form SEE level for each task. Our study expands on the Berry et al. study in that it includes both facet-specific SEE (confidence across levels of difficulty), which is similar to the measure of SEE strength in the study by Berry et al., and SEE level (confidence across four facets).

Locke et al. (1984) included *magnitude* (meaning level) and strength of SEE as two dimensions in their instrument to measure individuals' expectations about their ability to name uses for common objects. Much like the instrument developed by Berry et al. (1989), tasks at progressively more difficult levels were specified. Unlike Berry et al., Locke et al. (1984) investigated certainty ratings for varying levels of task difficulty, using strength of SEE for items related to similar levels (across tasks). The authors demonstrated that their instrument could differentiate between different levels of tasks, and found that SEE for moderate to difficult goal levels best predicted performance. They further argued that some previous goal-setting studies had been unsuccessful in detecting a positive relationship between expectancy beliefs and performance, due to a failure to account for different levels of performance. Reanalysing the previous studies, Locke et al. (1984) found positive and significant correlations between expectancy beliefs and performance outcomes for students in the medium and difficult goal-level groups, consistent with the results from their 1984 study.

Another example is the study by Zimmerman and Martinez-Pons (1990), who presented students with mathematics problems of increasing difficulty (from simple arithmetic to statistics) and asked them to rate their confidence (from 0 to 100%) to solve each problem. The measures used by Locke et al. (1984) and Zimmerman and Martinez-Pons (1990) both included varying levels of difficulty. However, while Locke et al. investigated the effect of levels of task difficulty directly, the latter study combined the scores from different levels of difficulty into a composite measure, similar to pure SEE strength as described by Lee and Bobko (1994) (see section 2.2.).

We found only one study relating to mathematics SEE that includes level of difficulty in the results. Chen and Zimmerman (2007) included three levels of difficulty in their

analyses when investigating the relationship between levels of task difficulty and students' SEE in mathematics in the US and Taiwan. Students' SEE for solving problems were analysed together with students' subsequent performance in solving the problems. Unlike the findings from the earlier studies by Locke et al. (1984), Chen and Zimmerman (2007) found that level of task difficulty was negatively related to students' accuracy in predicting their own task performance. While Locke et al. (1984) specified varying levels of task demand a priori, the mathematics problems in Chen and Zimmerman (2007) were placed in easy, moderate, or difficult categories as determined by post-hoc assessment of students' performance. Given that SEE are related to individuals' perceptions of capability, it is not certain that the easy, moderate, or difficult task levels in the study by Chen and Zimmerman (2007) are aligned with how the participants perceived the difficulty of the tasks before solving them. Thus the contradictory findings by Chen and Zimmerman (2007) and Locke et al. (1984) are not related to the same variables, and are therefore hard to compare.

In the instrument we propose in this study, level of SEE is related to students' perceptions of difficulty of future facets of taking a mathematics test. Levels of task difficulty are specified a priori, and included in the measurement, similarly to the procedures used by Berry et al. (1989) and Locke et al. (1984). Level of SEE includes students' certainty rating (strength of SEE) for each item related to the specific level, across facets. This is similar to the conception of level of SEE used by Locke et al. (1984).

2.4. A model proposal

Given the lack of instruments that include different levels of perceived difficulty when measuring mathematics SEE, this paper aims to fill the gap by proposing a multidimensional self-efficacy measure. In more detail, in this study we propose and investigate the structural validity of a new multidimensional instrument, which disentangles facets and levels from the strength of SEE for prospective national test performance. To this end we propose a theoretical model, which is presented in Figure 1. For the 14 items measuring strength of SEE we propose two factor structures to fit to the data:

- four test taking facets, i.e., latent constructs above indicators in Figure 1 depicting SEE to (1) complete a certain number of problems, (2) solve tasks of a certain challenge, (3) concentrate, and (4) not give up for a certain amount of time (Bandura 2006; Pajares 1996a; Zimmerman and Kitsansas 2005); and

- three perceived difficulty levels, i.e., latent constructs below indicators in Figure 1 depicting SEE to solve (1) easy, (2) medium difficulty, and (3) hard tasks of the test (Berry et al. 1989; Locke et al. 1984).

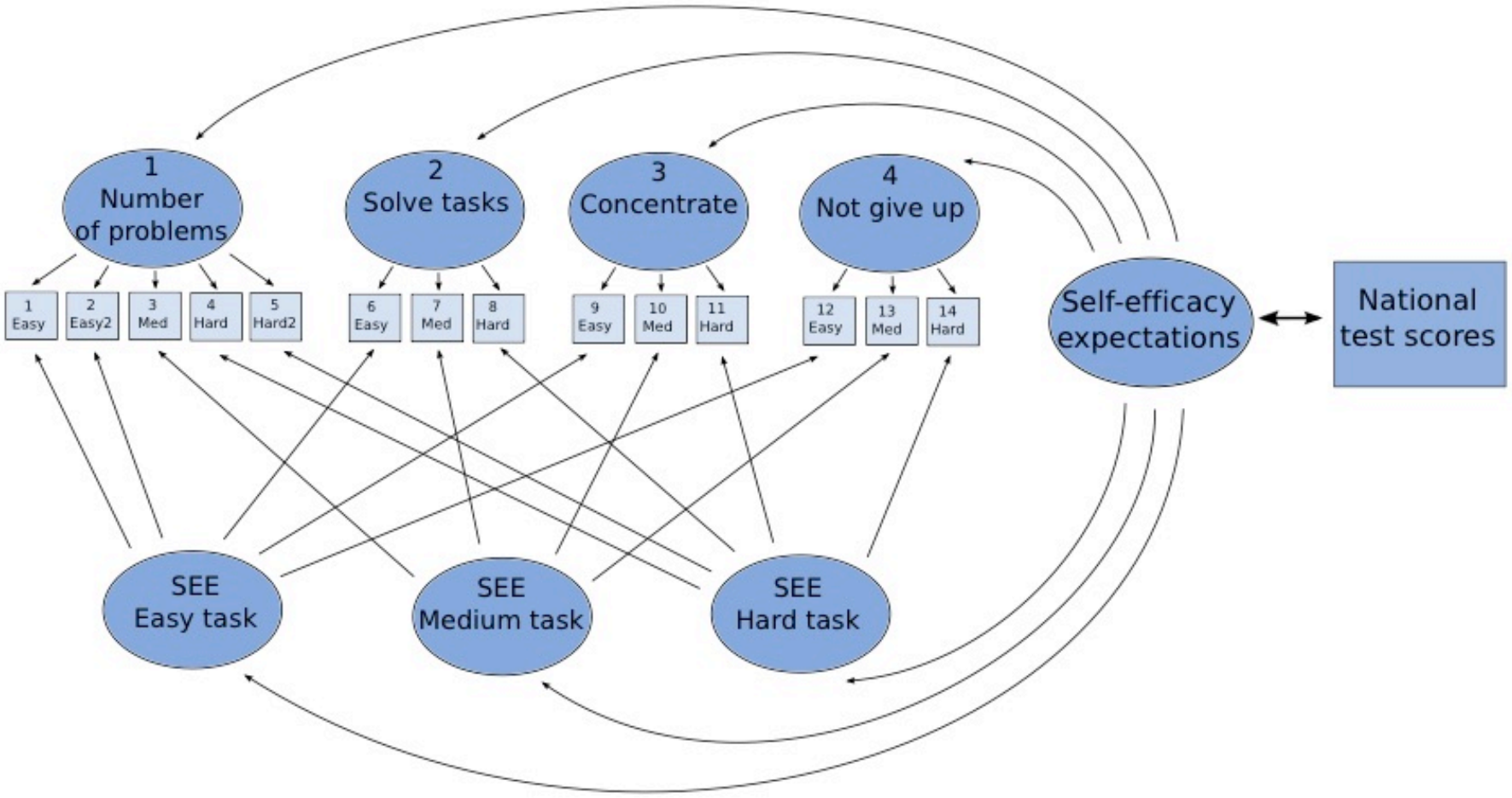
Subsequently, we hypothesise that facet-specific constructs are associated with students' national test performance, controlling for level. Likewise we hypothesise that level constructs are associated with students' national test performance, controlling for facets. We also hypothesize that strength could be a function of the combined scores from all 14 items, disregarding latent facets and levels, and thus an inferior estimate of SEE given the multidimensional nature of the present measure.

To conclude, the research questions we address in this paper are as follows:

(1) What is the structural validity of the proposed self-efficacy constructs, which include facets and levels of self-efficacy expectations?

(2) How are facet-specific and level-specific self-efficacy expectations associated with performance on national tests in mathematics?

Figure 1: Theoretical model



Note: SEE = Self-Efficacy Expectations
Items 1-14 are measured on a scale from 0 "not at all certain" to 10 "completely certain", representing strength of self-efficacy.
The four latent constructs above the indicators are test taking facets, while the three latent constructs below the indicators are levels of perceived difficulty.

3. Methodology

3.1 Sample

Participants were 756 students (364 female) in grades 5, 8, and 9 (10, 13 and 14 years old, respectively) in Norwegian state schools, which were part of a larger study (see Sørlie and Söderlund 2015). Participating schools were selected from municipalities within four regions in Norway, which differ in their socioeconomic characteristics. In Norway more than 96% of students attend state schools (Statistics Norway 2015) with little or no forms of tracking or ability grouping. A total of 38 classes from 27 schools participated, and 84% of the students asked agreed to participate in the study. The data collected consisted of questionnaires measuring students' SEE and scores from national tests in mathematics. To assess face-validity, the instrument was piloted with six 4th grade students of varying mathematics proficiency.

3.2. Instruments

3.2.1. Self-efficacy

Students' SEE were assessed with the Self-Efficacy Gradations of Difficulty Questionnaire (Appendix A), developed for the present study. This new instrument includes 14 items, where each item is related to one of four facets[2] concerning problem solving and self-regulation on national tests in mathematics, as well as one of three levels of difficulty within each facet. For each item participants were asked to indicate their confidence on an 11-point scale from 0 "not at all certain" to 10 "highly certain", which represents strength of SEE.

When we inspected descriptives (see Table 1) only items 1, 2, 6 and 9 were skewed and peaked (Kline 2011, p. 63). As most students are likely to be able to complete the easy tasks, it is reasonable that items associated with this level are non-normally distributed.

3.2.2. National test scores

National test scores in numeracy (Norwegian Directorate for Education 2016) were used as the performance measure and dependent variable. What is referred to as "numeracy" is essentially the same as what researchers frequently refer to as mathematics. Students in 5th grade are presented with 45 problems while students in 8th and 9th grades sit the same test, with 58 problems to solve. Each problem is scored as either 1 (correct) or 0 (incorrect). To enable meaningful analyses, we used students' absolute scores on the test (out of 45 or 58

[2] A 5th facet, "Grade", was originally tested, but not deemed adequate. Examination of individual student responses indicated that the items were confusing.

possible) rather than their resulting grade, which is dependent upon the national average each year. National test raw scores were linked to the questionnaire data by unique identifiers.

Table 1

Item descriptives

		Grade 5						Grade 8/9					
	Measure	n	M	SD	Range	Skew	Kurt	n	M	SD	Range	Skew	Kurt
1	No. of problems_very easy	337	10.68	1.34	10	-5.01*	27.37*	371	10.72	1.19	10	-5.35*	31.08*
2	No. of problems _easy	337	9.98	2.01	10	-2.11	3.99	371	10.25	1.65	10	-2.74	8.12
3	No. of problems _med	337	8.38	2.73	10	-0.72	-0.57	372	8.85	2.42	10	-0.97	**0.14**
4	No. of problems _hard	337	6.13	3.07	10	**0.03**	-1.02	372	6.37	2.72	10	**-0.00**	-0.81
5	No. of problems _very hard	337	3.98	2.95	10	0.66	-0.65	370	3.73	2.93	10	0.88	**-0.34**
6	Solve tasks_easy	336	10.13	1.67	10	-2.85	9.51	371	10.11	1.53	9	-2.40	6.29
7	Solve tasks_med	337	7.78	2.10	10	**-0.17**	-0.59	372	8.01	2.10	10	-0.67	0.43
8	Solve tasks_hard	337	5.08	2.73	10	**0.17**	-0.85	371	5.27	2.60	10	**0.10**	-0.84
9	Concentrate _easy	337	10.29	2.08	10	-3.33*	10.59*	371	10.24	2.07	10	-3.03*	8.84
10	Concentrate _med	337	9.38	2.40	10	-1.57	1.86	371	9.41	2.29	10	-1.56	1.83
11	Concentrate _hard	337	7.66	3.08	10	-0.63	-0.73	372	7.40	3.08	10	-0.60	-0.69
12	Not give up _easy	337	9.24	2.59	10	-1.46	1.42	371	9.24	2.56	10	-1.53	1.59
13	Not give up _med	337	8.16	2.76	10	-0.63	-0.71	372	8.26	2.54	10	-0.83	**0.14**
14	Not give up _hard	337	7.00	3.38	10	-0.35	-1.24	371	6.78	3.00	10	-0.23	-0.92
15	National test score	351	27.62	8.61	40	-0.28	-0.50	372	30.83	11.63	53	-0.05	-0.75

Note: z-score is larger than 1.65 for all items, *except* those in **bold**

*Absolute value higher than recommended cut-off

3.3. Empirical models

In order to test multidimensionality of our measure, we specified a set of logical models, ranging from simple (e.g., correlated latent constructs) to complex (e.g., correlated uniquenesses, higher-order constructs) (see Appendix B). In our theoretical model (Model 7), latent constructs represent both levels and facets, based on the logic of the multi-trait multi-method (MTMM) model (Campbell and Fiske 1959). Given that these types of models are especially prone to non-convergence or improper solutions (Marsh et al. 2010), we also specified models with correlated uniquenesses (CUs) representing levels or facets, an alternative strategy in the MTMM logic. While models with CUs do not provide estimates of the size of effects (as do models with latent constructs), both strategies enable us to consider both facets and levels of mathematics SEE.

In our baseline model (Model 1) all 14 items were specified to load on a single self-efficacy construct, representing overall strength of SEE. We then proceeded to specify models in which the items loaded on four facet constructs (Model 2), and three level constructs (Model 3). We specified in the "facet-model" a priori CUs for three levels (Model 2b), while in the "level-model" we specified CUs for four facets (Model 3b). We then specified higher order models in which an underlying self-efficacy construct predicts four facets (Model 4), and three levels (Model 5). Again, we specified for the "facet-model" a priori CUs for three levels (Model 4b), and for the "level-model" a priori CUs for four facets (Model 5b).

Finally we tested two models that include latent constructs to capture levels and facets. In Model 6, we specified four facets and three levels as latent, correlated factors, presumed independent of one-another. In Model 7, our theoretical model, we specified self-efficacy as a higher order construct predicting four latent facet constructs and three latent level constructs.

After testing the fit of our specified models, we investigated associations between students' SEE and national test scores. In doing so, we used a full structural equation model including national test scores as the correlate.

3.4. Analyses

The data were analysed with structural equation modelling using the Mplus software (version 7.31 for Mac: Muthén and Muthén 2012). As some items were non-normally distributed we used the robust maximum likelihood estimator. Data analyses were carried out separately for grade 5 and grades 8/9. We first inspected the correlation matrices (Table 2), and then carried out confirmatory factor analyses.

As indices of good model fit we used the chi square (χ^2/df =<3 acceptable), the Root Mean Square Error of Approximation (RMSEA <.08 acceptable), the Standardized Root

Mean Square Residual (SRMR<.10 acceptable), the Comparative Fit Index (CFI) and the Tucker-Lewis index (TLI) (CFI/TLI > .90 acceptable) (Chiorri 2014; Morin et al. 2014; Schermelleh-Engel et al. 2003). Models with "acceptable" values on a minimum of three of the five indices were considered.

After identifying models with acceptable fit, these were compared in regard to improvement in fit relative to model parsimony. The following cut-offs were used: ΔRMSEA (.015), ΔCFI (.010), ΔSRMR (.030). The use of fit indices should not replace sound theoretical judgment, and many researchers advocate including substantive and practical considerations when selecting the best models (Chiorri 2014; Schermelleh-Engel et al. 2003). The retained models were compared in regard to whether the parameter estimates were meaningful with reasonable standard errors, in order to arrive at a best-fit model.

MTMM models are particularly susceptible to non-convergence (Chiorri 2014). As many as 15% of properly specified models can result in non-positive residual variances (Chen et al. 2001). When we encountered negative residuals, step-by-step recommendations by Chen et al. (2001) were followed. More than one negative residual variance was taken as a sign of model misspecification, and any such models were rejected.

4. Results

4.1. Correlation coefficients

Correlation coefficients are given in Table 2. Some overall tendencies emerge across the 5th grade sample and the 8th/9th grade sample. Consistently, each item is most strongly correlated with the item closest to it in regard to level, within the same facet. For example, item 9 (concentrate_easy) is strongly correlated with item 10 (concentrate_med), but less so with item 11 (concentrate_hard). Thus, items are not always strongly correlated with other items within the same facet, if they are at opposite ends in regard to their respective associations with levels of difficulty.

In addition, there is a tendency for strong item correlations between levels of difficulty across facets. An example is item 8 (solve tasks_hard): within each facet it is most strongly correlated with the item(s) associated with the hard level, followed by items associated with the medium and the easy levels. Of the self-efficacy items, five cross-correlations are not significant in the grade 5 score, and two in the grade 8/9 scores. These cross-correlations are consistently associated with different levels of difficulty, warranting a MTMM type of model.

Table 2

Zero-order correlation coefficients among measures

	Measure	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	No. of problems_very easy		.76**	.45**	.21**	.01	.29**	.31**	.17**	.36**	.24**	.14*	.27**	.24**	.20**	.29**
2	No. of problems_easy	.61**		.70**	.46**	.20**	.44**	.51**	.35**	.43**	.41**	.34**	.38**	.37**	.30**	.37**
3	No. of problems _med	.32**	.70**		.74**	.41**	.45**	.62**	.51**	.30**	.41**	.44**	.40**	.48**	.42**	.43**
4	No. of problems _hard	.11	0.47**	.75**		.73**	.38**	.60**	.64**	.16**	.33**	.46**	.34**	.42**	.43**	.31**
5	No. of problems_very hard	.09	.33**	.60**	.77**		.25**	.46**	.56**	.03	.23**	.43**	.17**	.25**	.30**	.11*
6	Solve tasks_easy	.26*	.28**	.30**	.28**	.26**		.59**	.35**	.34**	.41**	.36**	.34**	.30**	.21**	.18**
7	Solve tasks_med	.24**	.40**	.53**	.56**	.48**	.37**		.76**	.30**	.42**	.50**	.39**	.49**	.45**	.36**
8	Solve tasks_hard	.15**	.35**	.51**	.56**	.56**	.17*	.72**		.15*	.30**	.50**	.29**	.42**	.49**	.29**
9	Concentrate _easy	.35**	.21*	.09	-.03	-.01	.21*	.16*	.15**		.67**	.24**	.44**	.37**	.22**	.34**
10	Concentrate _med	.24**	.36**	.38**	.31**	.24**	.22**	.48**	.44**	.56**		.64**	.47**	.49**	.39**	.31**
11	Concentrate _hard	.18**	.37**	.43**	.38**	.35**	.20**	.46**	.47**	.26**	.68**		.35**	.48**	.51**	.18**
12	Not give up _easy	.20*	.25**	.25**	.17**	.11*	.31**	.28**	.24**	.35**	.33**	.31**		.80**	.47**	.42**
13	Not give up _med	.28**	.38**	.46**	.39**	.29**	.27**	.44**	.43**	.28**	.43**	.44**	.71**		.79**	.44**
14	Not give up _hard	.23**	.37**	.50**	.42**	.35**	.15**	.45**	.48**	.23**	.40**	.45**	.50**	.87**		.35**
15	National test score	.11**	.27**	.32**	.35**	.28**	.17**	.38**	.38**	.12*	.17**	.26**	.27**	.32**	.25**	

Note: Grade 5 estimates are shown below the diagonal, grade 8/9 estimates are shown above the diagonal.

*Correlation is significant at the p<.05 level (2-tailed)

**Correlation is significant at the p<.01 level (2-tailed)

Estimates are from Mplus (Muthén and Muthén 2012)

4.2. Confirmatory factor analyses

The results from confirmatory factor analyses were largely similar for grade 5 and grade 8/9 scores (see Table 3), thus these results are presented together. For both samples, five models were initially accepted, based on fit indices. The baseline "unidimensional" model (Model 1) had very poor fit, as did Models 2, 3, 4, and 5. These models estimate fewer parameters than the five accepted models, which contributes to their worse fit. Models 2b, 3b, 4b, 5b and 6 were deemed "acceptable" in regard to fit indices. Model 7 did not converge. The best-fitting models in absolute terms are Models 3b and 5b (identical fit values), against which we compared the other models. In the grade 5 scores, Models 3b and 5b are substantially better fitting than Model 6, while the differences in fit values are generally smaller than the recommended cut-offs compared with Models 2b and 4b. In the grade 8/9 scores, Models 3b and 5b are substantially better fitting than all the alternative models. Given that Models 3b and 5b are more parsimonious and have better fit indices, Models 3b and 5b are preferred for both samples.

Besides fit indices we considered model parameters, with regard to their standard errors, residual variances and whether the estimates were meaningful in substantive terms. For both samples, Models 2b, 3b, 4b and 5b all result in largely meaningful parameter estimates. Model 6 gives several substantively problematic estimates, with nonsignificant factor loadings and large standard errors. All models except Model 3b result in one or several negative residual variances. In Models 2b and 4b, all items associated with the medium level factor have negative residual variance estimates, giving non-meaningful estimates for CUs between these items. Model 5b results in negative residual variance on the medium level factor, and was subjected to the tests recommended by Chen et al. (2001). These tests indicated the negative estimate was unlikely to be due to sampling fluctuations.

Table 3

Fit indices for confirmatory factor analyses

	Model	Grade 5						Grade 8/9					
		χ^2	p	RMSEA	CFI	TLI	SRMR	χ^2	p	RMSEA	CFI	TLI	SRMR
1	Unidimensional	$\chi^2_{(77)}$=882.49	p<.001	.176	.541	.458	.116	$\chi^2_{(77)}$=1343.40	p<.001	.210	.503	.413	.119
		χ^2/df=11.46						χ^2/df=17.45					
2	Correlated latent facets*	$\chi^2_{(71)}$= 407.28	p<.001	.119	.808	.754	.100	$\chi^2_{(71)}$=796.27	p<.001	.165	.716	.636	.120
		χ^2/df=5.74						χ^2/df=11.21					
2b	Correlated latent facets with	$\chi^2_{(45)}$= 121.23	p<.001	.071	.957	.912	.069	$\chi^2_{(45)}$= 150.40	p<.001	.079	.959	.916	.084
	CUs for 3 levels**	χ^2/df=2.69						χ^2/df=3.34					
3	Correlated latent levels	$\chi^2_{(74)}$= 810.69	p<.001	.172	.580	.484	.109	$\chi^2_{(74)}$=1622.57	p<.001	.237	.393	.253	.114
		χ^2/df=10.96						χ^2/df=21.93					
3b	Correlated latent levels with	$\chi^2_{(55)}$= 115.18	p<.001	.057	.966	.943	.057	$\chi^2_{(55)}$= 99.07	p<.001	.046	.983	.971	.052
	CUs for 4 facets	χ^2/df=2.09						χ^2/df=1.80					
4	2nd-order latent facets**	$\chi^2_{(73)}$= 416.73	p<.001	.118	.804	.756	.104	$\chi^2_{(73)}$=804.65	p<.001	.164	.713	.642	.129
		χ^2/df=5.71						χ^2/df=11.02					
4b	2nd-order latent facets with	$\chi^2_{(47)}$= 124.88	p<.001	.070	.956	.914	.073	$\chi^2_{(47)}$= 168.48	p<.001	.083	.952	.908	.089
	CUs for 3 levels**	χ^2/df=2.66						χ^2/df=3.58					
5	2nd-order latent levels*	$\chi^2_{(74)}$= 810.68	p<.001	.172	.580	.484	.109	$\chi^2_{(74)}$=1622.55	p<.001	.237	.393	.253	.114
		χ^2/df=10.96						χ^2/df=21.93					
5b	2nd-order latent levels with	$\chi^2_{(55)}$= 115.18	p<.001	.057	.966	.943	.057	$\chi^2_{(55)}$= 99.07	p<.001	.046	.983	.971	.052
	CUs for 4 facets ***	χ^2/df=2.09						χ^2/df=1.80					
6	Latent constructs model **	$\chi^2_{(54)}$= 158.62	p<.001	.076	.940	.900	.069	$\chi^2_{(54)}$= 196.99	p<.001	.084	.944	.906	.065
		χ^2/df=2.94						χ^2/df=3.65					
7	2nd-order latent constructs	Did not converge						Did not converge					
	model by facets and levels												

*Negative residual variance on 1 item **Negative residual variance on 2 or more items ***Negative residual variance on 1 factor

Overall, these results indicate that the best-fitting model when considering fit values is also the best-fitting in regard to parameter estimates (substantive interpretations, standard errors, and residual variances). Models 3b and 5b result in identical fit indices, however Model 5b has negative residual variance on the medium level factor, and was rejected. In both samples, Model 3b was retained as the best-fitting model, as it results in the best model fit indices, meaningful parameter estimates, and no negative residual variances. In this model, latent correlations are .82 (for grade 5) and .89 (for grades 8/9) between the easy and medium levels, .97 (for grade 5) and .90 (for grades 8/9) between the medium and hard levels, and .68 (for both grade 5 and grades 8/9) between the easy and hard levels.

4.3. Associations between national test results and self-efficacy

Having selected a best-fitting measurement model (Model 3b), we used this as the measurement part of a full structural equation model including national test scores as the correlate. To avoid issues of co-linearity due to the high correlation between the medium and hard level constructs (see section 4.2.), national test scores were correlated with, rather than regressed on, the three level constructs. We also specified as full structural equation models the unidimensional model (Model 1), for comparison purposes, and our theoretical model (Model 7). Model 7 does not converge as measurement model alone, but when test scores are included it does.

Results from the structural equation models are presented in Table 4. For both samples, a structural equation model based on the unidimensional measurement model (Model 1) results in poor fit, while Models 3b and 7 result in acceptable fit indices. Differences in fit between models are larger in the older student sample, however fit indices for both samples clearly indicate Model 3b as best-fitting, as expected.

Table 4

Fit indices for structural equation models

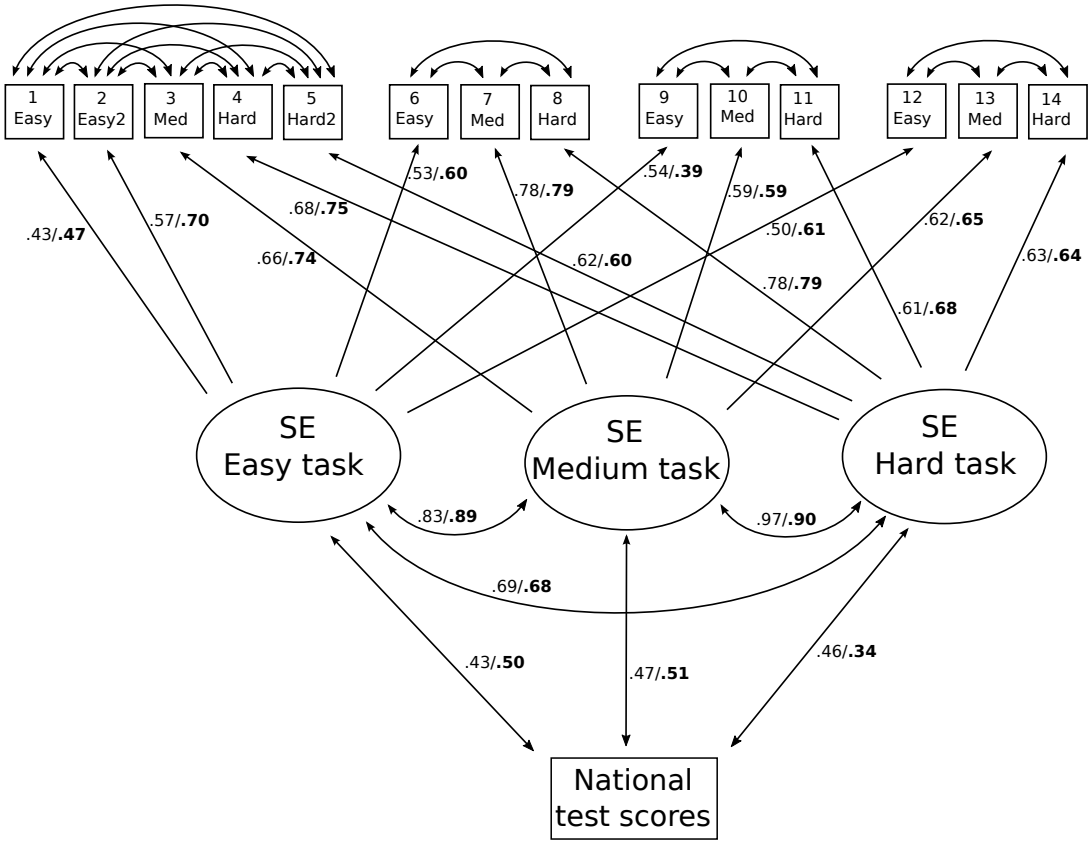
	Model	Grade 5						Grade 8/9					
		χ^2	p	RMSEA	CFI	TLI	SRMR	χ^2	p	RMSEA	CFI	TLI	SRMR
1	Test scores correlated w/ unidimensional model	$\chi^2_{(90)}$= 946.27 χ^2/df= 10.51	p<.001	.162	.548	.473	.110	$\chi^2_{(90)}$=1424.55 χ^2/df= 15.83	p<.001	.194	.511	.429	.114
3b	Test scores correlated w/ correlated levels w/ CUs for facets	$\chi^2_{(66)}$= 140.80 χ^2/df=2.13	p<.001	.056	.961	.937	.056	$\chi^2_{(66)}$= 139.65 χ^2/df=2.12	p<.001	.053	.973	.957	.056
7	Test scores correlated w/ 2nd order latent constructs model by facets and levels**	$\chi^2_{(69)}$= 174.71 χ^2/df=2.53	p<.001	.065	.944	.915	.061	$\chi^2_{(69)}$= 226.87 χ^2/df=3.29	p<.001	.076	.942	.912	.062

**Negative residual variance on 2 or more items

Again, we considered model parameters. Models 1 and 3b both result in meaningful factor loadings, small standard errors, and no negative residual variances in relation to both samples. In Model 1, the correlation between overall SEE strength and performance scores is .46 for grade 5, and .48 for grades 8/9. In contrast, the hypothesised theoretical model (Model 7) results in non-significant factor loadings and/or large standard errors for most items on three out of seven first order factors, as well as for the level factors on the self-efficacy construct. Also, there are negative residual variances on all items associated with the medium level factor, as well as the medium level construct itself.

When considering both fit indices and parameter estimates, Model 3b is retained as the best-fitting structural equation model. In this model the highest correlation estimate for both samples is between national test scores and the medium level factor (grade 5: .47, grades 8/9: .51). With grade 5, the second strongest correlation is between test scores and the hard level factor (.46), followed by the easy level factor (.43). In contrast, with grades 8/9 the estimate for the easy level factor (.50) indicates a stronger relationship with national test scores than the hard level factor (.38). Factor loadings from Model 3b are presented in Figure 2. This indicates that students' scores on national tests are related to level of SEE. Stronger SEE are related to higher national test results, and this relationship is strongest for SEE related to tasks of medium levels of perceived difficulty.

Figure 2: Structural equation model
(Model 3b)



Note: Estimates before "/" are Grade 5 results, after "/", in **bold**, are grade 8/9 results.
All estimates are standardized (STYX) estimates from Mplus (Muthén & Muthén 2012).
Estimates for CUs are not given due to space limitations.

5. Discussion

In this study, we investigated the structural validity of the Self-Efficacy Gradations of Difficulty Questionnaire, which disentangles test taking facets and levels of difficulty from strength of SEE, and how the aforementioned constructs were associated with performance on national tests in mathematics. Next we discuss our findings, organised by the research questions.

5.1. RQ1: Structural validity of our proposed measure

According to the most common conceptualisation of self-efficacy (Bandura 1997; Schunk and Pajares 2010; Zimmerman 2000), SEE vary according to specificity (facets), level (levels of difficulty), and strength (confidence). The present findings indicate that it is reasonable to measure mathematics SEE with a multidimensional instrument, including latent constructs representing test taking facets and levels of difficulty. Models with a single SEE strength indicator, or including facet- or level-constructs without CUs, fitted the data less well. Hence, multidimensionality needs to be accounted for in the models over and above unidimensionality.

In the present study, the best-fitting measurement model is a three correlated levels model with CUs specified according to four facets (Model 3b). This indicates that the students in our study differentiated between levels of perceived difficulty of test taking facets when forming their SEE. The fact that this finding was replicated for two separate samples strengthens the generalizability of the result. Previous research has emphasised the importance of task-specificity of mathematics SEE (Bong 2001; Pajares and Miller 1995). The importance of including levels of difficulty in self-efficacy assessment has been asserted (Bandura 2006; Pajares 1996b; Zimmerman 2000) and demonstrated empirically (Berry et al. 1989; Locke et al. 1984), although it does not seem to have previously been tested in mathematics. Our study demonstrates the relevance of considering levels of difficulty when measuring mathematics SEE.

The best fitting model has better fit indices in the case of the older (grades 8/9) than the younger (grade 5) students, and differences in model fit between best and second best models were larger with the grade 8/9 sample. Also, the high correlation between medium and hard level tasks indicated grade 5 students did not discriminate between these. These results might all reflect that older students take more account of different levels of difficulty than younger students, supporting previous research indicating that students increasingly differentiate between motivational constructs with age (Bong 2001). An implication for classroom practice is that students can benefit from working on tasks of varying difficulty levels, as they develop their SEE. Further studies are needed into the relationship between student age and SEE for different levels of perceived difficulty.

In the current study, each item was related both to a facet and to a level of difficulty, and for each item, strength of SEE was measured on an 11-point scale. To the best of our knowledge SEE in mathematics has not previously been measured in this way. Further replications are needed to confirm the validity of this approach.

5.2. RQ2: Multidimensional self-efficacy expectations and national test scores in mathematics

Structural equation models were specified on the basis of the measurement models tested for in RQ1. We proposed a theoretical model (see Figure 2), which was operationalised in a second order latent constructs measurement model including latent facets and latent levels (Model 7). This model did not converge as measurement model for the present samples and resulted in non-meaningful parameter estimates when specified as structural equation model. It is rare to encounter models with more than one negative residual variance, and recommendations in this regard are inconclusive. While the model did not converge to a proper solution for the present samples this is not evidence the model is "wrong". This model should be tested in future studies, as it could provide information about the relationship with performance scores for facets and levels of SEE.

In the current samples, a MTMM model in which students' scores on a national test in mathematics are related to three latent factors representing SEE for different levels of difficulty fit data best (Model 3b). CUs were specified according to four test taking facets. The present findings indicate there is a relationship between level of students' SEE and their scores on national tests in mathematics. In both grade 5 and grades 8/9, stronger SEE are related to higher national test scores. This result is in support of previous research (e.g., Multon et al. 1991; Pajares 1996b; Stajkovic and Sommer 2000). It appears that the relationship between SEE and national test scores is moderated by age, where older students' SEE are more strongly related to their national test scores than younger students' SEE. This result is also in line with previous research findings (Bong 2001).

Chen and Zimmerman (2007) suggested task difficulty might influence the accuracy of students' SEE. In our study, the relationship between SEE and national test scores is strongest for SEE related to medium levels of perceived difficulty. Our current study thus supports the findings from Locke et al. (1984), as opposed to those from Chen and Zimmerman (2007). Locke et al. (1984) decided to include data from only the medium and hard levels of difficulty in the analyses (excluding the easy and extremely hard levels), justified because of the higher association with performance scores for SEE at these levels. We do not advocate this practice, given that it is not yet known whether the medium and hard levels of difficulty are consistently most strongly associated with performance, or whether this relationship differs for different facets of mathematics. This relationship may also differ according to student age, as exemplified in our findings. In grade 5 hard tasks had the second strongest correlation of national test scores, while for grades 8/9 easy tasks had the second

strongest correlate. This result may possibly be related to grade 8/9 students' test experiences in grade 5, or to the increased difficulty of the test in grade 8/9. Further studies are needed into the relationship between levels of SEE and performance outcomes for different facets of mathematics, and for different age groups.

Another possible avenue of research is to link levels of perceived difficulty with performance on different level tasks. In our study we included level of difficulty in the SEE measure a priori, while another study determined level of difficulty post hoc according to student performance (Chen and Zimmerman, 2007). The predictive utility of our questionnaire might have increased if we had linked SEE level with performance scores on items of corresponding levels of difficulty.

6. Limitations

First, our findings are limited to the levels and facets included in our measure, which was designed to be representative but not exhaustive. Future studies should consider including levels of perceived difficulty for additional facets of mathematics, such as learning new mathematical ideas with differing degrees of help from the teacher. Second, with the sample we used some theoretically plausible models led to non-convergence and improper solutions. Perhaps some items were too easy, indicated by the skewness of some easy level items. It would be valuable to carry out a future simulation study to investigate some likely reasons for this. Third, the findings are limited to Norwegian students in grades 5, 8 and 9. It is possible that a future replication of the study in a different socio-cultural context could provide new insights.

7. Conclusion

We tested a new instrument responding to the proposed multidimensionality of SEE, including levels of perceived difficulty as well as facets of a national test. The results indicated that students differentiate between levels of perceived difficulty, as well as test taking facets, when forming their SEE. Further, we tested the associations between SEE as measured by our instrument, and national test scores. We found that levels of SEE were meaningfully related to national test scores, and that the correlation with test scores was highest for SEE in relation to tasks of medium level of perceived difficulty.

Findings from the present study suggest that researchers should consider including levels of perceived difficulty when measuring mathematics SEE. Students' SEE may vary not only as a function of specific tasks or facets, but according to the perceived difficulty of these tasks. Furthermore, what kind of task is perceived to be easy or hard differs between students, and students' perceptions of difficulty may differ from their actual performance and from the

level of difficulty as determined by others. To further our understanding of the relationship between task difficulty and students' SEE, future studies can investigate the relationships between mathematical tasks as subjectively perceived or experienced by students, and as 'objectively' determined by teachers or researchers.

Teachers of mathematics may wish to consider levels of task difficulty as experienced by students, to support the formation of positive SEE by regulating the difficulty level of tasks students work on. In our study we found that students differentiated more between easy and medium level tasks, than between medium and hard level tasks. We propose it is valuable to the formation of accurate SEE that students be encouraged to attempt difficult tasks with appropriate levels of support, and be given individual feedback on their performance.

Acknowledgments:
The Self-Efficacy Gradations of Difficulty Questionnaire was created as part of a study funded by the Norwegian Research Council (Norges Forskningsråd); the grant number is 218282/H20 (PRAKUT: Learning regions project). We wish to thank Göran Söderlund and Anita Nordheim, collaborators on the Learning regions study, for conceptual contributions towards this instrument.

References

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: W. H. Freeman & Co.

Bandura, A. (2006). Guide for constructing self-efficacy scales. In F. Pajares & T. C. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (pp. 307–337). Greenwich: Information Age Publishing.

Berry, J. M., West, R. L., and Dennehey, D. M. (1989). Reliability and validity of the Memory Self-Efficacy Questionnaire. *Developmental Psychology*, *25*(5), 701–713.

Bong, M. (2001). Between- and within-domain relations of academic motivation among middle and high school students: Self-efficacy, task-value, and achievement goals. *Journal of Educational Psychology*, *93*(1), 23–34.

Bong, M. (2009). Age-related differences in achievement goal differentiation. *Journal of Educational Psychology*, *101*(4), 879–896.

Bong, M., and Skaalvik, E. M. (2003). Academic self-concept and self-efficacy: How different are they really? *Educational Psychology Review*, *15*, 1–40.

Campbell, D. T., and Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105.

Chen, F., Bollen, K. A., Paxton, P., Curran, P. J., & Kirby, J. B. (2001). Improper solutions in Structural equation models. *Sociological Methods & Research*, *29*(4), 468–508.

Chen, P., and Zimmerman, B. J. (2007). A cross-national comparison study on the accuracy of self-efficacy beliefs of middle-school mathematics students. *Journal of Experimental Education*, *75*(3), 221–244.

Chiorri, C. (2014). Competing Factor Structures for the Toronto Empathy Questionnaire. In D. Wyatt and J. Panksepp (Eds.), *Psychology of empathy* (pp. 1–61). New York: Nova Science Publishers.

Dowling, D. M. (1978). *The development of a mathematics confidence scale and its application in the study of confidence in women college students*. Ohio State University.

Hackett, G., and Betz, N. E. (1989). An exploration of the mathematics self-efficacy/ mathematics performance correspondence. *Journal for Research in Mathematics Education*, *20*(3), 261–273.

Kline, R. B. (2011). *Principles and practice of Structural Equation Modeling* (3rd ed.). New York: The Guildford Press.

Lee, C., and Bobko, P. (1994). Self-efficacy beliefs: Comparison of five measures. *Journal of Applied Psychology*, *79*(3), 364–369.

Locke, E. A., Frederick, E., Lee, C., and Bobko, P. (1984). Effect of self-efficacy, goals, and task strategies on task performance. *Journal of Applied Psychology*, *69*(2), 241–251.

Marsh, H. W., Scalas, L. F., and Nagengast, B. (2010). Longitudinal tests of competing factor structures for the Rosenberg Self-Esteem Scale: Traits, ephemeral artifacts, and stable response styles. *Psychological Assessment*, *22*(2), 366–381.

Martin, A. J., Anderson, J., Bobis, J., Way, J., and Vellar, R. (2012). Switching on and switching off in mathematics: An ecological study of future intent and disengagement among middle school students. *Journal of Educational Psychology*, *104*(1), 1–18.

Meece, J. L., Wigfield, A., and Eccles, J. S. (1990). Predictors of math anxiety and its influence on young adolescents' course enrollment intentions and performance in mathematics. *Journal of Educational Psychology*, *82*(1), 60–70.

Morin, A. J. S., Marsh, H. W., Nagengast, B., and Scalas, L. F. (2014). Doubly Latent Multilevel Analyses of Classroom Climate: An Illustration. *The Journal of Experimental Education*, *82*(2), 143–167.

Muis, K. R., and Duffy, M. C. (2013). Epistemic climate and epistemic change: Instruction designed to change students' beliefs and learning strategies and improve achievement. *Journal of Educational Psychology*, *105*(1), 213–225.
Multon, K. D., Brown, S. D., and Lent, R. W. (1991). Relation of self-efficacy beliefs to academic outcomes: A meta-analytic investigation. *Journal of Counseling Psychology*, *38*(1), 30–38.
Muthén, L. K., and Muthén, B. O. (2012). *Mplus User`s Guide* (7th ed.). Los Angeles, CA: Muthén and Muthén.
Norwegian Directorate for Education. (2016). Nasjonale prøver. http://www.udir.no/Vurdering/Nasjonale-prover/ Accessed 28.04.2016
Pajares, F. (1996a). Self-Efficacy Beliefs and Mathematical Problem-Solving of Gifted Students. *Contemporary Educational Psychology*, *21*(4), 325–44.
Pajares, F. (1996b). Self-efficacy beliefs in academic settings. *Review of Educational Research*, *66*(4), 543–578.
Pajares, F., and Miller, M. D. (1995). Mathematics self-efficacy and mathematics performances: The need for specificity of assessment. *Journal of Counseling Psychology*, *42*(2), 190–198.
Phan, H. P. (2013). Self-Efficacy, Reflection, and Achievement: A Short-Term Longitudinal Examination. *The Journal of Educational Research*, *00*, 1–13.
Pintrich, P. R., and de Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology*, *82*(1), 33–40.
Schermelleh-Engel, K., Moosbrugger, H., and Müller, H. (2003). Evaluating the Fit of Structural Equation Models : Tests of Significance and Descriptive Goodness-of-Fit Measures. *Methods of Psychological Research Online*, *8*(2), 23–74.
Scholz, U., Dona, B. G., Sud, S., and Schwarzer, R. (2002). Is general self-efficacy a universal construct? Psychometric findings from 25 countries. *European Journal of Psychological Assessment*, *18*(3), 242–251.
Schukajlow, S., Leiss, D., Pekrun, R., Blum, W., Müller, M., and Messner, R. (2012). Teaching methods for modelling problems and students' task-specific enjoyment, value, interest and self-efficacy. *Educational Studies in Mathematics*, *79*(2), 215–237.
Schulz, W. (2005). Mathematics Self-Efficacy and Student Expectations. Results from PISA 2003. In *Annual meetings of the American Educational Research Association* (pp. 11–15). Montreal.
Schunk, D. H., & Pajares, F. (2010). Self-Efficacy Beliefs. In P. Peterson, E. L. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education* (3rd ed., pp. 668–672). New York: Elsevier Ltd.
Schweinle, A., and Mims, G. A. (2009). Mathematics self-efficacy: stereotype threat versus resilience. *Social Psychology of Education*, *12*(4), 501–514.
Stajkovic, A. D., and Sommer, S. M. (2000). Self-Efficacy and Causal Attributions : Direct and Reciprocal Links. *Journal of Applied Social Psychology*, *30*(4), 707–737.
Statistics Norway. (2015). *Elevar i grunnskolen, 1.oktober 2015*. https://www.ssb.no/utdanning/statistikker/utgrs/aar/2015-12-11#content Accessed 11.04.2016
Sørlie, K., and Söderlund, G. (2015). Regionale forskjeller i mestringsforventninger. In G. Langfeldt (Ed.), *Skolens kvalitet skapes lokalt. Presentasjon av funn fra forskningsprosjektet "Lærende Regioner."* Oslo: Fagbokforlaget.
Valentine, J. C., Dubois, D. L., and Cooper, H. M. (2004). The Relation Between Self-Beliefs and Academic Achievement: A Meta-Analytic Review. *Educational Psychologist*,

39(2), 111–133.
Zeldin, A. L., and Pajares, F. (2000). Against the Odds: Self-Efficacy Beliefs of Women in Mathematical, Scientific, and Technological Careers. *American Educational Research Journal*, *37*(1), 215–246.
Zimmerman, B. J. (2000). Self-Efficacy: An Essential Motive to Learn. *Contemporary Educational Psychology*, *25*(1), 82–91.
Zimmerman, B. J., and Kitsantas, A. (2005). Homework practices and academic achievement: The mediating role of self-efficacy and perceived responsibility beliefs. *Contemporary Educational Psychology*, *30*(4), 397–417.
Zimmerman, B. J., and Martinez-Pons, M. (1990). Student differences in self-regulated learning: Relating grade, sex, and giftedness to self-efficacy and strategy use. *Journal of Educational Psychology*, *82*(1), 51–59.

Appendix B: Conference proceedings

Self-efficacy and mathematics performance: reciprocal relationships

Karin Elisabeth Sørlie Street[1], Lars-Erik Malmberg[2] and Gabriel J. Stylianides[3]
[1] Oxford University, Department of Education, Oxford, United Kingdom *and* Høgskulen i Sogn & Fjordane, Sogndal, Norway; karin.street@stx.ox.ac.uk
[2, 3] Oxford University, Department of Education, Oxford, United Kingdom;
[2]lars-erik.malmberg@education.ox.ac.uk [3]gabriel.stylianides@education.ox.ac.uk

Several studies have looked at either the effects of students' Self-Efficacy Expectations (SEE) on their mathematics performance results, or the effect of previous mathematics performances on students' SEE. Few studies have tested the theoretically proposed reciprocal relationship between mathematics SEE and performance in mathematics. Furthermore, previous studies have not included levels of difficulty, although this is an integral part of the definition of SEE. This study applied a new measure of SEE, which included both test taking facets and levels of perceived difficulty, to investigate their relationships with students' performance on a national test in mathematics. Two models gave very good fit to the data, and supported the reciprocal effects model. These models provided estimates of the relationships between students' test performance and their level and facet-specific SEE, respectively.

Keywords: Self-efficacy, performance, reciprocal effects, levels of difficulty, test taking facets.

Introduction

Self-efficacy expectations (SEE) are important because extensive research indicates they are related to student learning and performance (Zimmerman, 2000). Studies have demonstrated the effect of mathematics SEE on performance results (e.g. Pajares, 1996) and vice versa (e.g. Pampaka, Kleanthous, Hutcheson, & Wake, 2011), but the proposed reciprocal relationship between SEE and performance (Bandura, 1997) has received little attention. Investigating this relationship empirically is important in order to understand the relationship between SEE and performance in mathematics, and the process by which SEE may influence performance in mathematics and vice versa. Few studies have investigated reciprocal effects between SEE and mathematics performance with longitudinal data, and none that we know of in relation to national tests in Norway.

Bandura (1997) argues SEE vary according to three dimensions, but we know of no studies that have included SEE *level of difficulty* to investigate reciprocal effects with mathematics. As we argued in Street, Malmberg, & Stylianides (in review), including level of SEE is important as it is an integral part of the definition of SEE (Bandura, 1997). We aimed to address this research gap by testing a reciprocal effects model including students' level and facet-specific SEE, and their scores on a Norwegian national test. To this aim, we applied a recently developed measure of mathematics SEE that includes four facets of test taking as well as three levels of perceived difficulty. We included students' SEE responses and test scores as they progressed from grade 8 to grade 9.

Theoretical background

Self-efficacy expectations (SEE) are judgments about individuals' perceived capability to perform on future tasks (Bandura, 1997), for instance students' expectations they are able to carry out the various facets involved with taking a mathematics test. Examples are problem-solving skills such as solving a certain number of problems or solving tasks of a certain

challenge, or skills in self-regulation such as concentrating for a length of time or persevering through difficult problems.

SEE affect future performance through mediating processes, influencing students' behaviors and motivations. SEE may influence individuals' tendencies to approach learning tasks, their effort and persistence while engaged in such tasks, as well as their self-regulatory processes (Zimmerman, 2000). SEE have been found to predict mathematical problem-solving after controlling for factors such as cognitive ability, mathematics Grade Point Average, anxiety, and gender (Pajares, 1996). There is also empirical support for the influence of past performances on SEE (e.g. Pampaka et al). According to Bandura (1997), SEE are formed through four sources, where mastery experiences is the strongest source. Mastery experiences stem from individuals' appraisals of previous performance situations, for instance their experiences from a previous, similar test. Importantly, previous experiences do not influence SEE directly, but are interpreted and made sense of by the individual (Usher, 2009).

A mutually reinforcing pattern of influence between self-beliefs and academic achievement is supported for several types of self-beliefs (Valentine, Dubois, & Cooper, 2004). Studies have, for instance, investigated reciprocal effects between self-concept and achievement (Marsh, Trautwein, Lüdtke, Köller, & Baumert, 2013), however only few studies have investigated reciprocal effects between SEE and mathematics performance. Williams and Williams (2010) proposed a reciprocal effects model, which fit the data well for PISA results from 26 out of 33 countries. They argued reciprocal effects might be a fundamental psychological process, however the data from Norway was excluded from the analyses due to little variation in the grade-level variable. Hannula et al. (Hannula, Bofah, Tuohilampi, & Metsämuuronen, 2014) investigated the dominant direction of effect between mathematics performance and SEE, as students progressed from grade 3 to grade 9 in Finland. Their results supported a reciprocal effects model, where the effect from performance on SEE (.30) was somewhat stronger than the effect from SEE on performance (.26). Both measures were also relatively stable over time. The authors noted the effect from SEE on performance was larger for the older students in the sample.

Measures of SEE have typically included SEE strength and SEE specificity (also called generality). While Bandura proposes a third dimension of SEE (1997), only few studies have included level of difficulty. Students' SEE may vary between different facets of mathematics, and in regards to different perceived levels of difficulty. Locke et al. (Locke, Frederick, Lee, & Bobko, 1984) investigated SEE in relation to the task naming uses for common objects, and found that SEE for tasks of medium to high difficulty were most predictive of subsequent task performance.

Theoretical model and research questions

In line with theory and empirical findings discussed, we propose a theoretical model (see Figure 1), where SEE and mathematics performance are reciprocally related. We included Norwegian students' SEE responses and subsequent test scores from the beginning of grades 8 and 9 to investigate this model. In the model, students' SEE in grade 8 and 9 are related to their performance in mathematics in the same year. Furthermore, students' mathematics performances in grade 8 predict their SEE in grade 9. Both SEE and mathematics performance are relatively stable (mathematics performance the most stable of the two), with the grade 8 constructs predicting scores in grade 9. We further propose different SEE constructs (level and facet-specific SEE) have differential relationships with mathematics performance.

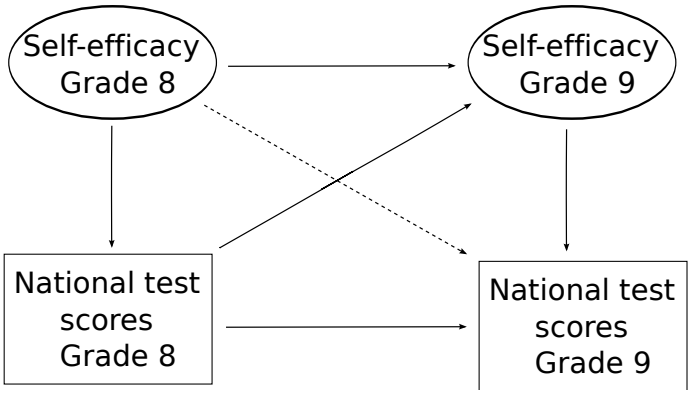


Figure 1: Theoretical model

RQ1: What is the relationship between SEE and performance on national tests in mathematics in grade 8, and SEE and performance on national tests in mathematics in grade 9?
RQ2: Do the relationships between self-efficacy expectations and national test scores differ according to levels or facets of SEE?

Methodology

Participants

The participants were 95 students (44 female) in Norwegian secondary school who completed self-report questionnaires and took national tests in mathematics at the beginning of grades 8 and 9 (13 and 14 years old). The participants were part of a larger sample, selected for cross-sectional investigations (see Sørlie & Söderlund, 2015). Included in the study were schools where students had performed above and below what might be expected on national tests, considering measures of socio-economic-status. For a detailed explanation of this strategy see Langfelt (2015).

Measures

The Self-efficacy Gradations of Difficulty Questionnaire was applied to measure SEE at the two time points (see Street et al., in review, for a detailed analysis). This is a recently developed multidimensional measure of mathematics SEE, that includes four test taking facets (facet-specific SEE) related to problem solving (complete a number of problems, solve tasks of a certain challenge) and self-regulation (concentrate, not give up), and three levels (easy, medium, and hard) of perceived difficulty (level of SEE). Each of the 14 items in the measure are related to one test taking facet and one level of difficulty within each facet (see Figure 2 for an example facet, ”concentrate”, with three levels of difficulty). For each item, students are asked to indicate their confidence (strength of SEE) on an 11-point scale from 0 “not at all certain” to 10 “highly certain”. The structural validity of this measure was tested (Street et al., in review), and the resulting best-fit measurement model included three latent (unobserved) level constructs, with correlated uniquenesses (correlated error terms) specified for each of the four facets.

Some people can find it hard to concentrate when sitting tests. How certain are you that you will be able to concentrate during the national tests in mathematics?

During the national tests this year I will be able to concentrate:

	0 Not at all certain	1	2	3	4	5 Moderately certain	6	7	8	9	10 Completely certain
at least a little while											
at least half the time											
the whole time											

Figure 2: Example from the Self-Efficacy Gradations of Difficulty Questionnaire

The performance measure was raw scores from national tests in numeracy (Norwegian Directorate for Education, 2016). "Numeracy" is similar to what researchers generally refer to as mathematics. Norwegian students sit these tests at the start of the school year in grades 5, 8 and 9. Students in grades 8 and 9 sit the same test, which involves 58 problems, scored as either correct (1) or incorrect (0). We used unique identifiers to link the national test scores with the questionnaire responses.

Specification of empirical models

Our modeling choices were informed by the previously established factor structure of our measure, as well as our theoretical model. Street et al. (in review) found that the best-fit models were those that accounted for the multidimensional nature of SEE, through correlated latent constructs and correlated uniquenesses. In order to estimate the relationships between latent constructs in our models we used item parceling technique. Parceling involves aggregating (taking the sum or average) two or more items to manufacture an indicator of a construct (Little, Cunningham, Shahar, & Widaman, 2002, p. 152). While our measure of SEE is multidimensional, the factor structure has been tested and established in a previous study. Item parcels were formed to achieve "clean" latent constructs from each year, through aggregating items in such a way that the secondary loading was spread across parcels (Little et al., 2002). For example, the "easy" parcel contains all items related to the easy latent construct, across four facets. Similarly, the "concentrate" parcel includes all items related to the concentrate facet, across three levels of difficulty. As an example, the "concentrate" parcel is created through summing the scores of the three items included in Figure 2.

Two reciprocal effects models were specified. Our hypothesized model (Model 1) includes national test scores and item parcels for SEE levels, representing the best-fit model from the previous study. The alternative model (Model 2) includes national test scores and item parcels for facet-specific SEE. In both models, SEE constructs in grade 8 are related to test scores in grade 9. Similarly, SEE constructs in grade 9 are related to test scores in grade 9. SEE constructs in grade 8 are intercorrelated, as are SEE constructs in grade 9. Furthermore, corresponding SEE constructs are related across the two years (e.g. SEE easy in grade 8 predicts SEE easy in grade 9). Test scores in grade 8 predict test scores in grade 9. Finally, the reciprocal relationships tested are the path from test scores in grade 8 on SEE constructs in grade 9, and the paths from SEE constructs in grade 8 on test scores in grade 9.

Analyses

We analysed the data with structural equation modeling, using the maximum likelihood estimator in Mplus (version 7.31 for Mac: Muthén & Muthén, 2012). We used fit indices recommended by previous studies (e.g. Morin, Marsh, Nagengast, & Scalas, 2014; Schermelleh-Engel, Moosbrugger, & Müller, 2003), specifically we used the chi square (χ^2/df =<3 acceptable), the Root Mean Square Error of Approximation (RMSEA <.08 acceptable), the Standardized Root Mean Square Residual (SRMR<.10 acceptable), the Comparative Fit Index (CFI) and the Tucker-Lewis index (TLI) (CFI/TLI > .90 acceptable). To assess

improvement in fit between models, we used the following cut-offs: ΔRMSEA (.015), ΔCFI (.010), ΔSRMR (.030).

Results

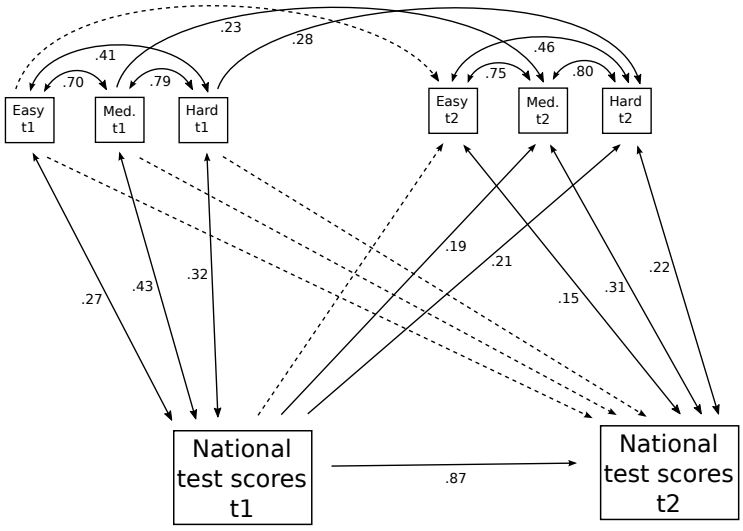
Item cross-correlations for items from the same year were consistent with previous results from the same measure. The correlation matrix indicates items are related to two types of latent constructs; levels (easy, medium, and hard) and facets (no. of problems, solve tasks, concentrate, and not give up). Autocorrelations between items in grade 8 and the corresponding (same) item in grade 9 demonstrate two tendencies. First, autocorrelations are significant for all items, except for two cases, both associated with the easy level (no. of problems_2, and concentrate_easy). Second, the magnitudes of the associations are consistently weaker for items associated with the easy level, than the medium and hard levels. To illustrate this: the strongest autocorrelation of the easy level items is .21, while the weakest autocorrelations of the medium and hard level items is .26 and .29, respectively. Thus, the cross-correlations for each of the years indicate support for a multidimensional construct, while the autocorrelations indicate that students' scores on the medium and hard level items were more stable across time, than students' scores on the easy level items. The non-significant autocorrelations between the grade 8 and grade 9 easy level items might be related to a lack of variability in scores. Most students were highly confident in relation to the easy level items.

Fit indices for confirmatory factor analyses

	Model	χ^2	*p*	RMSEA	CFI	TLI	SRMR
1	Correlated levels model, reciprocal relationship with test scores	$\chi^2_{(6)}$= 4.397 χ^2/df=0.73	*p*=.623	.000	1.00	1.02	.048
2	Correlated facets model, reciprocal relationship with test scores	$\chi^2_{(12)}$= 10.890 χ^2/df=0.91	*p*=.538	.000	1.00	1.01	.061

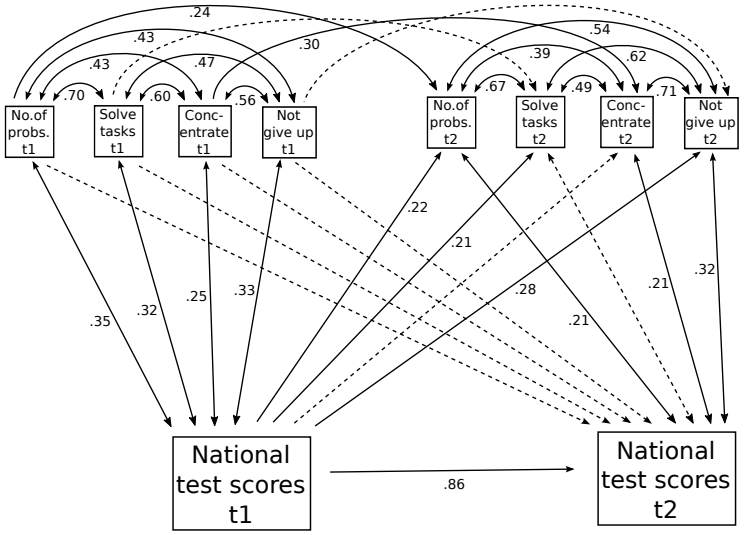
Table 1: Results from structural equation models

Results from the structural equation models are presented in Table 1. Both our proposed models resulted in excellent fit. The aim of the present study is not model comparison, but to test a reciprocal effects model, as well as to estimate the relationships between national test scores and level and facet-specific SEE, respectively. Accordingly, parameter estimates from both models are reported (see Figures 3 and 4).



Note: "Easy", "Med.", and "Hard" refer to easy, medium and hard levels of perceived difficulty. "t1" and "t2" refer to time 1 and time 2.
All estimates are standardized (STYX) estimates from Mplus (Muthén & Muthén 2012).
Paths are significant at the $p<.05$ level. Dotted paths are nonsignificant.

Figure 3: Structural equation model, Model 1 (correlated levels model)



Note: "No. of probs.", "Solve tasks", "Concentrate", and "Not give up" refer to facets of test taking. "t1" and "t2" refer to time 1 and time 2.
All estimates are standardized (STYX) estimates from Mplus (Muthén & Muthén 2012).
Paths are significant at the p<.05 level. Dotted paths are nonsignificant.

Figure 4: Structural equation model, Model 2 (correlated facets model)

As can be seen in Figure 3, results from Model 1 partially support a reciprocal effects model between levels of SEE and national test scores. Specifically, national test scores in grade 8 predict levels of SEE in grade 9 in relation to medium and hard, but not easy, tasks. The predicted relationship from SEE level constructs in grade 8 to national test scores in grade 9 is not supported. However, all three levels of SEE are associated with test scores the same year, including when previous test performance is controlled in grade 9. Two further relationships are nonsignificant: the path from SEE easy level in grade 8 to grade 9, and the path from national test scores in grade 8 to SEE easy level in grade 9.

Model 2 results (see Figure 4) are generally consistent with results for Model 1. Test scores in grade 8 predict three facets of SEE in grade 9 (not "concentrate"), while the paths from SEE facet constructs in grade 8 to test scores in grade 9 are nonsignificant. However, facet-specific self-efficacy constructs are related to performance results in each year, including in grade 9 (not "solve tasks") when previous test scores are included in the model. Unlike the other facets, the magnitude of the relationship between "not give up" and test scores is very similar in grade 8 and 9. Similar to the easy level in Model 1, the relationships between "solve tasks" and "not give up" from grade 8 to grade 9 are not significant. Finally, we see that test scores are more stable over time than either level or facet-specific SEE.

Discussion

The current study lent empirical support to a reciprocal relationship between level and facet-specific SEE, and performance in mathematics (RQ1). It was demonstrated that students' SEE predicted scores on a test in the immediate future, while national test scores predicted their SEE one year later. All SEE items in our study are explicitly linked with the immediately upcoming test, thus it is reasonable the SEE constructs are associated with performance results of the immediately upcoming tests, but not with the tests one year later. In two previous studies, measures of SEE and performance were collected over the course of a single day (Locke et al., 1984) or three years apart (Hannula et al., 2014). Findings from both these studies indicated the dominant effect was from performance to SEE. The type of test, and whether it is considered important by the students, might influence the relationship between performance experiences and SEE. National tests are quite "talked about" in school in Norway, and students in grades 8 and 9 are likely becoming increasingly aware of upcoming exams (at the end of year 10). This might have provided quite a memorable context for their test experiences.

Furthermore (RQ2), differential relationships were found between national tests and easy, medium and hard levels of perceived difficulty. The associations with test scores were stronger for medium and hard levels of difficulty, which is in line with previous research (Locke et al., 1984). The relationships with test performance also varied between different facets of test taking, where "not give up" provided the strongest association.

The findings indicate it may be useful for students to work on tasks of medium level difficulty, to develop accurate SEE as well as skills in mathematics. Furthermore, they indicate that as well as problem solving SEE, SEE for self-regulation is related to students' subsequent test performance. This implies classroom time might be well spent developing self-regulatory as well as problem solving skills.

The current findings are limited in that we could not control for other factors that may have influenced students' SEE and test performance. Also, while the measure of SEE provided interesting information, partitioning the effects of test taking facets and levels of perceived difficulty, further replications are needed with a larger sample, and in different cultural contexts.

References

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: W. H. Freeman and Co.

Hannula, M. S., Bofah, E. A., Tuohilampi, L., & Metsämuuronen, J. (2014). A longitudinal analysis of the relationship between mathematics-related affect and achievement in Finland. *Proceedings of the Joint Meeting of PME 38 and PME-NA 36*, *38*, 249–256.

Langfeldt, G. (2015). Skolens kvalitet skapes lokalt. In *Skolens kvalitet skapes lokalt. Presentasjon av funn fra forskningsprosjektet "Lærende Regioner."* Oslo: Fagbokforlaget.

Little, T. D., Cunningham, W. a., Shahar, G., & Widaman, K. F. (2002). To Parcel or Not to

Parcel: Exploring the Question, Weighing the Merits. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 151–173.
Locke, E. A., Frederick, E., Lee, C., & Bobko, P. (1984). Effect of self-efficacy, goals, and task strategies on task performance. *Journal of Applied Psychology*, *69*(2), 241–251.
Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2013). Academic self-concept, interest, grades, and standardized test scores: reciprocal effects models of causal ordering. *Child Development*, *76*(2), 397–416.
Morin, A. J. S., Marsh, H. W., Nagengast, B., & Scalas, L. F. (2014). Doubly Latent Multilevel Analyses of Classroom Climate: An Illustration. *The Journal of Experimental Education*, *82*(2), 143–167.
Muthén, L. K., & Muthén, B. O. (2012). *Mplus User`s Guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.
Norwegian Directorate for Education. (2016). Nasjonale prøver. Retrieved April 28, 2016, from http://www.udir.no/Vurdering/Nasjonale-prover/
Pajares, F. (1996). Self-Efficacy Beliefs and Mathematical Problem-Solving of Gifted Students. *Contemporary Educational Psychology*, *21*(4), 325–44.
Pampaka, M., Kleanthous, I., Hutcheson, G. D., & Wake, G. (2011). Measuring mathematics self-efficacy as a learning outcome. *Research in Mathematics Education*, *13*(2), 169–190.
Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the Fit of Structural Equation Models : Tests of Significance and Descriptive Goodness-of-Fit Measures. *Methods of Psychological Research Online*, *8*(2), 23–74.
Sørlie, K., & Söderlund, G. (2015). Regionale forskjeller i mestringsforventninger. In G. Langfeldt (Ed.), *Skolens kvalitet skapes lokalt. Presentasjon av funn fra forskningsprosjektet "Lærende Regioner."* Oslo: Fagbokforlaget.
Usher, E. L. (2009). Sources of Middle School Students ' Self-Efficacy in Mathematics : A Qualitative Investigation. *American Educational Research Journal*, *46*(1), 275–314.
Valentine, J. C., Dubois, D. L., & Cooper, H. M. (2004). The Relation Between Self-Beliefs and Academic Achievement: A Meta-Analytic Review. *Educational Psychologist*, *39*(2), 111–133.
Williams, T., & Williams, K. (2010). Self-efficacy and performance in mathematics: Reciprocal determinism in 33 nations. *Journal of Educational Psychology*, *102*(2), 453–466.
Zimmerman, B. J. (2000). Self-Efficacy: An Essential Motive to Learn. *Contemporary Educational Psychology*, *25*(1), 82–91.

Appendix C: Differences between chapter 2 in my thesis and a published article, and a comment on authorship.

Findings from some of the research questions in chapter 2 have been published in Street, Malmberg, and Stylianides (2017; see Appendix A). However, the study that is presented in chapter 2 of the thesis is different from the published article in some important ways.

First, the chapter is longer and more in-depth as compared to the published article. An indication of this is the difference in length of the documents: whereas the main body of the article consists of 7.572 words, chapter 2 in the thesis consists of 29.750 words. The article includes two figures and four tables (plus an overview of empirical models in an attachment), while the current chapter includes 15 figures and 23 tables (including the overview of the empirical models). These differences in length are a reflection of the chapter being more in-depth, offering the kind of information one would expect to see in a doctoral thesis but would not necessarily appear in a published article where there is a strong pressure on space. For instance in the thesis chapter I am able to go into more detail when discussing previous measures of self-efficacy. The thesis chapter also contains a section discussing gender differences in self-efficacy, which is not included in the published article. As further examples, the thesis chapter offers more methodological details, such as an expanded discussion about ethical considerations, background information about the development of my measure, and detailed descriptions about the empirical models I tested. The thesis chapter also provides a more extensive discussion of implications.

Second, the chapter differs substantively from the published article because it includes investigation of a further research question. The additional research question *(What is the relationship between students' self-efficacy expectations for perceived easy, medium, and hard tasks, and their scores on easy, medium, and hard level items on national tests in mathematics?)* enables investigations into the relationship between self-efficacy expectations and national test scores for items of matching levels of difficulty, and was not included in the published article. The results pertaining to this additional research question are treated in section 2.4.4, and discussed in section 2.5.3, neither of which were included in the published article. In conclusion, although the chapter is based on the same dataset and investigates some of the same themes as the journal article, it differs significantly in format, length, and substantive content.

Comments on authorship
Given that the published article is written by myself and co-authored by my DPhil supervisors, I will briefly comment on the distribution of work and related authorship. For my DPhil chapter (this applies to the whole thesis, of course) as well as the journal article, I have designed the studies, made the final decisions on research methodology and analyses, I conducted all the data collection, entered the data and carried out data quality control, and I wrote the actual text. During these activities, my supervisors have supported my development through helping me narrow down the particular focus of the article/chapter and situate it in a way that would make clear its contribution to existing research, asking questions, critiquing my work/writing/ideas, and providing hints and suggestions; in relation to the journal article, my supervisors also helped me navigate the review process, which was new to me as a beginning researcher. In short, they have supervised me.

Taber (2012) discusses the issue of DPhil students publishing together with their supervisors, and whether "normal" supervision work could and should qualify the supervisors as co-

authors of students' manuscripts. According to Taber (2012, p.6) *An author of an academic work is someone who made a substantial intellectual contribution to the work.* He further argues (p.7) *in general if a research supervisor is not making substantive intellectual input into the development of the student project we might wonder if they are fulfilling their role as a supervisor*. Finally, Taber comments on the input by supervisors to a thesis submitted as monograph (p.7) *the thesis is a stand-alone monograph that is presented as 'all the student's own work'. In the latter case, it is well-understood that the student is supported by a supervisor or advisory team, but that the sole authorship of the thesis reflects how the final research decisions, the actual collection and analysis of data and the drafting and honing of the thesis is undertaken by the student.* In the case of my thesis work, as well as the chapter in question, the work is "all mine", as is the normal requirement of a DPhil thesis. However, this does not mean my work has not benefitted from substantial intellectual contributions from my supervisors, to the degree that they qualify to be named as co-authors in the related journal article.

Another thing to consider are the ethical guidelines of my secondary affiliation, the University of Western Norway for applied sciences (previously the University College of Sogn & Fjordane). As is the case with many higher education institutions in Norway, the university follows the 'Vancouver protocol' (1997). In the section about authorship (p.4) it is stated *Authorship credit should be based only on substantial contributions to 1) conception and design, or analysis and interpretation of data; and to 2) drafting the article or revising it critically for important intellectual content; and on 3) final approval of the version to be published. Conditions 1, 2, and 3 must all be met*. While the requirements stated in the Vancouver protocol perhaps seem more stringent than the arguments made by Taber, the underlying principle is the same, in regard to substantial intellectual contributions. According to the arguments set forth both by Taber (2012) and in the Vancouver Protocol (1997), the contributions my supervisors made to the journal article are clearly worthy of co-authorship. At the same time, their input did not exceed that of active and involved supervisors, meaning the work on the thesis is "all mine", in accordance with the normal understanding of a supervised thesis.

References:

Case, K., Clever, L. H. Colaianni, L. A., Davidoff, F., Glass, R., Horton, R., .. Woolf, P. (1997). Uniform requirements for manuscripts submitted to biomedical journals. *Annals of Internal Medicine, 126*(1), 36-47.

Taber, K. S. (2012). Who counts as an author when reporting educational research? *Chemistry Education Research and Practice, 14*(1), 5-8.

Appendix D: Participant information letter, students

To the student/legal guardian Sogndal 21.09.2012

What

This letter is regarding participation in the research study "Efficacy beliefs". The study is part of a larger research project, "Learning regions", which is concerned with how we might facilitate teaching in schools to improve student attainment. The study will run for several years; autumn 2012 we invite 5th and 8th grade students in four different regions in Norway to participate, through answering a questionnaire. The 8th grade students will be invited to participate in a similar questionnaire autumn 2013. Some schools will be contacted later on, and invited to parttake in the next stage of the study, which involves classroom observations.

How

The questionnaire is either online or on paper (each school decides), and will be filled out at school, just before the national tests in mathematics this year. It will take about 20 minutes to complete. The questionnaire will be about the students` thoughts about the national tests, about mathematics, and about their own learning in school. Responses to the questionnaires will be analyzed in relation to the students` results on the national tests in mathematics.

The questionnaire information on-line is password protected, only the researchers have access to it, and it will be deleted from the web once the data is collected. National test results will be collected from the Norwegian directorate for Education (Udir), in accordance with their guidelines. Personal identification and questionnaire responses/national test results will be kept separate, and your child`s name, responses and/or test results will not be appear anywhere. Personal information will be kept in a secure locker, and will be deleted at the end of the project period.

Participation in this project is voluntary, and participants may at any time withdraw from the study withouth specifying a reason. The study has obtained approval from the Data protection official for research, NSD, and the University College of Sogn & Fjordane is responsible for the project.

Why

The aims of this study is to increase our understanding of what influences student attainment on national tests, how to facilitate student learning in schools, and to share this knowledge with other researchers, with school owners, principals and teachers in school. We hope results from this study might help schools and teachers facilitate learning in school for all students.

The results from this study will be published in one or several research reports of national and international kind, and will be used to share knowledge regarding teaching and learning with teachers and in schools.

What do I need to to?

1) If you consent to your child`s participation; please complete the below consent form, and hand it in at your child`s school
2) During autumn 2012 your child will complete a questionnaire at school, which takes about 20 minutes to complete (8th grade students will be invited to repeat this autumn 2013)
3) Some classes will be contacted later on regarding participation in the next stage of the research, which involves classroom observations

Kind regards,

Karin Sørlie
Principal investigator and University College lecturer, the University College of Sogn & Fjordane
Phone: 57676158
E-mail: karin.elisabeth.sorlie@hisf.no

Göran Søderlund
PhD and University College lecturer, the University College of Sogn & Fjordane
Phone: 57676069
E-mail: goran.soderlund@hisf.no

_____ (cut here) _____

Declaration of consent:

As the legal guardian of (name of student) …….. I have received written information, I consent to my child`s participation in the research study ”Efficacy beliefs”, and I consent to the study obtaining my child`s results from the national tests in mathematics, and linking these to his/her questionnaire responses.

Signature …….

Appendix E: Participant information letter, teachers /schools

Sogndal 28.08.2012

To teachers in 5th/8th grade

Background
This letter is regarding participation in the research study "Efficacy beliefs". The study is part of a larger research project, "Learning regions", where the University of Tromsø, the University of Agder, the University College of Nord-Trøndelag, the University college of Hedmark and the University College of Sogn & Fjordane are all participating. The project is concerned with how we might facilitate teaching in schools to improve student attainment. The study will run for several years; autumn 2012 we invite 5th and 8th grade students in four different regions in Norway to participate, through answering a questionnaire. The 8th grade students will be invited to participate in a similar questionnaire autumn 2013. Some schools will be contacted later on, and invited to parttake in the next stage of the study, which involves classroom observations.
The questionnaire will be filled out at school, just before the national tests in mathematics this year, and will take about 20 minutes to complete. It will be about the students` thoughts about the national tests, about mathematics, and about their own learning in school. Responses to the questionnaires will be analyzed in relation to the students` results on the national tests in mathematics.

Aims
The aims of this study is to increase our understanding of what influences student attainment on national tests, how to facilitate student learning in schools, and to share this knowledge with other researchers, with school owners, principals and teachers in school.

Ethical concerns
The results from this study will be published in one or several research reports of national and international kind, and will be used to share knowledge regarding teaching and learning with teachers, school owners and other researchers.
All personal information from the questionnaires and otherwise will be treated confidentially. No participants will be identifiable through the research reports published.
After the data collection the information will be stored safely, and personal information will be kept separate to questionnaire responses / national test results. The personal information will be deleted after a maximum of 5 years (20.09.2017). Information from this study, excluding personal information, might be shared with other researchers within the "Learning regions" project.
Please contact the principal investigator if you wish to see the student questionnaire.
Participation in this project is voluntary, and participants may at any time withdraw from the study withouth specifying a reason. The study has obtained approval from the Data protection official for research, NSD, and the University College of Sogn & Fjordane is responsible for the project.

What this entails:

1) If you wish to participate please return the below consent form to the administration at your school
2) As the teacher you will distribute and collect information letters/consent forms to the students in your class/ their legal guardians
3) You will administer the student questionnaire (paper or online) and be present when they fill it out during class
4) Some classes will be contacted later on to be asked if they wish to participate in the next stage of the research, involving classroom observations.

Kind regards,
Karin Sørlie
Principal investigator and University College lecturer, the University College of Sogn & Fjordane
Phone: 57676158
E-mail: karin.elisabeth.sorlie@hisf.no
Göran Søderlund
PhD and University College lecturer, the University College of Sogn & Fjordane
Phone: 57676069
E-mail: goran.soderlund@hisf.no

_____ (cut here) _____

Declaration of consent:

I have received written information, and consent to participating in the research study ”Efficacy beliefs”.

Signature …….

Appendix F: NSD approval & CUREC approval

Norsk samfunnsvitenskapelig datatjeneste AS
NORWEGIAN SOCIAL SCIENCE DATA SERVICES

NSD

Harald Hårfagres gate 29
N-5007 Bergen
Norway
Tel: +47-55 58 21 17
Fax: +47-55 58 96 50
nsd@nsd.uib.no
www.nsd.uib.no
Org.nr. 985 321 884

Karin Sørlie
Avdeling for lærarutdanning og idrett
Høgskulen i Sogn og Fjordane
Postboks 133
6851 SOGNDAL

Vår dato: 31.08.2012 Vår ref:31267 / 3 / MAS Deres dato: Deres ref:

TILBAKEMELDING PÅ MELDING OM BEHANDLING AV PERSONOPPLYSNINGER

Vi viser til melding om behandling av personopplysninger, mottatt 29.08.2012. Meldingen gjelder prosjektet:

31267	*Mestringsforventninger og regionale forskjeller på nasjonale prøver*
Behandlingsansvarlig	*Høgskulen i Sogn og Fjordane, ved institusjonens øverste leder*
Daglig ansvarlig	*Karin Sørlie*

Personvernombudet har vurdert prosjektet og finner at behandlingen av personopplysninger er meldepliktig i henhold til personopplysningsloven § 31. Behandlingen tilfredsstiller kravene i personopplysningsloven.

Personvernombudets vurdering forutsetter at prosjektet gjennomføres i tråd med opplysningene gitt i meldeskjemaet, korrespondanse med ombudet, eventuelle kommentarer samt personopplysningsloven og helseregisterloven med forskrifter. Behandlingen av personopplysninger kan settes i gang.

Det gjøres oppmerksom på at det skal gis ny melding dersom behandlingen endres i forhold til de opplysninger som ligger til grunn for personvernombudets vurdering. Endringsmeldinger gis via et eget skjema, http://www.nsd.uib.no/personvern/forsk_stud/skjema.html. Det skal også gis melding etter tre år dersom prosjektet fortsatt pågår. Meldinger skal skje skriftlig til ombudet.

Personvernombudet har lagt ut opplysninger om prosjektet i en offentlig database, http://www.nsd.uib.no/personvern/prosjektoversikt.jsp.

Personvernombudet vil ved prosjektets avslutning, 20.09.2017, rette en henvendelse angående status for behandlingen av personopplysninger.

Vennlig hilsen

Vigdis Namtvedt Kvalheim

Mads Solberg

Mads Solberg tlf: 55 58 89 28
Vedlegg: Prosjektvurdering

Avdelingskontorer / District Offices
OSLO: NSD, Universitetet i Oslo, Postboks 1055 Blindern, 0316 Oslo. Tel: +47-22 85 52 11. nsd@uio.no
TRONDHEIM: NSD, Norges teknisk-naturvitenskapelige universitet, 7491 Trondheim. Tel: +47-73 59 19 07. kyrre.svarva@svt.ntnu.no

28 March 2014 16:03

Dear Karin,

Application Approval

Title: *Students' self-efficacy beliefs in the mathematics classroom*

The above application has been considered on behalf of the Departmental Research Ethics Committee (DREC) in accordance with the procedures laid down by the University for ethical approval of all research involving human participants. I am pleased to inform you that, on the basis of the information provided to DREC, the proposed research has been judged as meeting appropriate ethical standards, and accordingly, approval has been granted.

Should there be any subsequent changes to the project which raise ethical issues not covered in the original application you should submit details to research.office@education.ox.ac.uk for consideration.

Good luck with your DPhil work.

Yours sincerely,

Nikki Luke
Member of DREC

Nikki Luke
Research Officer
Rees Centre for Research in Fostering and Education
University of Oxford Department of Education
15 Norham Gardens
Oxford OX2 6PY

nikki.luke@education.ox.ac.uk
+44 (0)1865 274050
http://reescentre.education.ox.ac.uk

Appendix G: Item descriptives, including 5th facet, "Grade"

		Grade 5						Grade 8/9					
	Measure	n	M	SD	Range	Skew	Kurt	n	M	SD	Range	Skew	Kurt
1	No. of problems_easy	337	10.68	1.34	10	-5.01*	27.37*	371	10.72	1.19	10	-5.35*	31.08*
2	No. of problems _easy2	337	9.98	2.01	10	-2.11	3.99	371	10.25	1.65	10	-2.74	8.12
3	No. of problems _med	337	8.38	2.73	10	-0.72	-0.57	372	8.85	2.42	10	-0.97	**0.14**
4	No. of problems _hard	337	6.13	3.07	10	**0.03**	-1.02	372	6.37	2.72	10	**-0.00**	-0.81
5	No. of problems _hard2	337	3.98	2.95	10	0.66	-0.65	370	3.73	2.93	10	0.88	**-0.34**
6	Minimum grade_easy	335	9.47	2.80	10	-1.77	2.06	371	9.47	3.39	10	-1.91	1.83
7	Minimum grade_easy2			*NA*				371	9.42	3.15	10	-1.84	1.84
8	Minimum grade_med	335	8.05	2.52	10	-0.55	-.30	370	8.94	2.78	10	-1.36	0.94
9	Minimum grade_hard			*NA*				371	7.69	2.52	10	-0.53	-0.40
10	Minimum grade_hard2	336	5.81	2.68	10	**0.06**	-.57	371	5.54	2.76	10	**0.03**	-0.87
11	Solve tasks_easy	336	10.13	1.67	10	-2.85	9.51	371	10.11	1.53	9	-2.40	6.29
12	Solve tasks_med	337	7.78	2.10	10	**-0.17**	-0.59	372	8.01	2.10	10	-0.67	0.43
13	Solve tasks_hard	337	5.08	2.73	10	**0.17**	-0.85	371	5.27	2.60	10	**0.10**	-0.84
14	Concentrate _easy	337	10.29	2.08	10	-3.33*	10.59*	371	10.24	2.07	10	-3.03*	8.84
15	Concentrate _med	337	9.38	2.40	10	-1.57	1.86	371	9.41	2.29	10	-1.56	1.83
16	Concentrate _hard	337	7.66	3.08	10	-0.63	-0.73	372	7.40	3.08	10	-0.60	-0.69
17	Not give up _easy	337	9.24	2.59	10	-1.46	1.42	371	9.24	2.56	10	-1.53	1.59
18	Not give up _med	337	8.16	2.76	10	-0.63	-0.71	372	8.26	2.54	10	-0.83	**0.14**
19	Not give up _hard	337	7.00	3.38	10	-0.35	-1.24	371	6.78	3.00	10	-0.23	-0.92
20	National test score	351	27.62	8.61	40	-0.28	-0.50	372	30.83	11.63	53	-0.05	-0.75

Note: z-score is larger than 1.65 for all items, *except* those in bold

*Absolute value higher than recommended cut-off

Appendix H: Grade 8/9 correlation matrix, including 5th facet, "Grade"

Zero-order correlation coefficients among measures

	Measure	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	No. of problems_easy																			
2	No. of problems _easy2	0.761**																		
3	No. of problems _med	0.453**	0.703**																	
4	No. of problems _hard	0.205**	0.455**	0.739**																
5	No. of problems _hard2	0.014	0.196**	0.413**	0.729**															
6	Minimum grade_easy	0.058	0.147*	0.067	-0.024	-0.105														
7	Minimum grade_easy2	0.112*	0.216**	0.128*	0.020	-0.078	0.959**													
8	Minimum grade_med	0.222**	0.328**	0.333**	0.214**	0.036	0.608**	0.698**												
9	Minimum grade_hard	0.202**	0.385**	0.566**	0.517**	0.315**	0.098	0.171**	0.506**											
10	Minimum grade_hard2	0.132**	0.290**	0.480**	0.604**	0.540**	-0.065	-0.020	0.237**	0.665**										
11	Solve tasks_easy	0.289**	0.438**	0.446**	0.380**	0.249**	0.137*	0.160**	0.277**	0.432**	0.337**									
12	Solve tasks_med	0.311**	0.512**	0.617**	0.600**	0.456**	0.078	0.112	0.319**	0.577**	0.556**	0.589**								
13	Solve tasks_hard	0.169**	0.353**	0.513**	0.640**	0.559**	-0.044	-0.004	0.210**	0.521**	0.657**	0.345**	0.758**							
14	Concentrate _easy	0.358**	0.433**	0.296**	0.160**	0.025	0.386**	0.395**	0.379**	0.282**	0.144*	0.344**	0.298**	0.145*						
15	Concentrate _med	0.239**	0.412**	0.414**	0.331**	0.232**	0.229**	0.260**	0.320**	0.377**	0.290**	0.408**	0.420**	0.302**	0.674**					
16	Concentrate _hard	0.136*	0.344**	0.437**	0.475**	0.426**	-0.009	0.024	0.143*	0.375**	0.456**	0.357**	0.499**	0.503**	0.237**	0.637**				
17	Not give up _easy	0.270**	0.384**	0.400**	0.337**	0.168**	0.201**	0.231**	0.332**	0.357**	0.254**	0.337**	0.394**	0.285**	0.436**	0.467**	0.353**			
18	Not give up _med	0.244**	0.372**	0.478**	0.424**	0.246**	0.126*	0.153*	0.326**	0.406**	0.337**	0.298**	0.485**	0.421**	0.367**	0.491**	0.475**	0.801**		
19	Not give up _hard	0.202**	0.304**	0.417**	0.434**	0.300**	-0.004	0.011	0.194**	0.373**	0.407**	0.213**	0.453**	0.492**	0.215**	0.386**	0.507**	0.471**	0.788**	
20	National test score	0.294**	0.373**	0.436**	0.311**	0.113*	0.215**	0.249**	0.331**	0.362**	0.251**	0.183**	0.362**	0.292**	0.340**	0.308**	0.181**	0.423**	0.440**	0.350**

*Correlation is significant at the p<.05 level (2-tailed)

**Correlation is significant at the p<.01 level (2-tailed)

Appendix I: Self-Efficacy Gradations of Difficulty, test version

Participants are first given a practice task (see Bandura, 2006, p.320).
For each statement, they are asked to tick the box which indicates how certain they are, on a scale from 0 to 10, with the anchors 0= ”Not at all certain”, 5=”Moderately certain”, and 10= ”Highly certain”.

The following sentences are related to the national tests in mathematics this year. Choose the number which best describes how certain you are.

There are more than 50 problems on the national tests in mathematics. How certain are you that you can solve at least a certain number of them?
During the national tests this year I can solve:
1: at least 5 of the problems
2: at least 10 of the problems
3: at least 25 of the problems
4: at least 40 of the problems
5: all the problems

During the national tests in mathematics some tasks might be easy and some might be hard to solve. How certain are you that you can solve these tasks?
During the national tests this year I can solve:
6: all the easy tasks
7: all the medium tasks
8: all the hard tasks

Some people can find it hard to concentrate when sitting tests. How certain are you that you will be able to concentrate during the national tests in mathematics?
During the national tests this year I will be able to concentrate:
9: at least a little while
10: at least half the test
11: the whole test

It might be easy to give up when faced with difficult tasks during a mathematics test. How certain are you that you can manage *not to give up* when faced with difficult tasks during the national tests in mathematics this year?
During the national tests this year I will manage not to give up:
12: when I am faced with slightly difficult tasks
13: when I am faced with quite difficult tasks
14: when I am faced with very difficult tasks

Appendix J: Self-Efficacy Gradations of Difficulty, classroom version

Developed by Karin Street
The following items differ from the original instrument, but follow the same format, and are based on the same conceptual idea of including facets and levels of difficulty of mathematics. The practice task below is from Bandura (2006, p.320).

The following sentences are related to the next few lessons in mathematics. Choose the number which best describes how certain you are.

1: During the next few lessons in mathematics you will learn a new topic. How certain are you that you can learn this new topic with a certain amount of help from the teacher?

During the next few lessons I can learn a new topic:

	0 Not at all certain	1	2	3	4	5 Moderately certain	6	7	8	9	10 Completely certain
if I get a lot of help from the teacher											
if I get some help from the teacher											
without any help from the teacher at all											

2: When learning new topics in mathematics students often practice through solving problems. How certain are you that you can solve at least a certain number of problems?

During each of the next few lessons I can solve:

	0 Not at all certain	1	2	3	4	5 Moderately certain	6	7	8	9	10 Completely certain
at least some of the problems I attempt											
about half of the problems I attempt											
all of the problems I attempt											

3: When solving new problems in mathematics, some problems might seem easy and some might seem hard *to you*. How certain are you that you can solve these tasks?

During the next few lessons I can solve:

	0 Not at all certain	1	2	3	4	5 Moderately certain	6	7	8	9	10 Completely certain
all the easy tasks											
all the medium tasks											
all the hard tasks											

4: Sometimes it can be hard to concentrate when doing mathematics. How certain are you that you will be able to concentrate when solving problems during each of the next few lessons?

I will be able to concentrate:

	0 Not at all certain	1	2	3	4	5 Moderately certain	6	7	8	9	10 Completely certain
at least a little while											
at least half the time											
the whole time											

5: It might be easy to give up when faced with difficult, new tasks in mathematics. How certain are you that you can manage *not to give up* if faced with difficult problems during the next few lessons?

During the next few lessons I can manage *not to give up*:

	0 Not at all certain	1	2	3	4	5 Moderately certain	6	7	8	9	10 Completely certain
if I am faced with slightly difficult tasks											
if I am faced with medium difficulty tasks											
if I am faced with very difficult tasks											

Appendix K: Self-Efficacy Gradations of Difficulty (SEGD), short classroom version.

Developed by Karin Street. Words in <brackets> were changed to reflect the topic of each lesson.

For each of these questions, think about the lesson on <algebra> which is just about to start.

1: In the next lesson in mathematics you will learn about <algebra>. How certain are you that you can learn <algebra> with a certain amount of help from the teacher?
In the next lesson in maths I can learn <algebra>:

	0 Not at all certain	1	2	3	4	5 Moderately certain	6	7	8	9	10 Completely certain
if I get a lot of help from the teacher											
if I get some help from the teacher											
without any help from the teacher at all											

2: When solving <algebra> problems in mathematics, some problems might seem easy and some might seem hard *to you*. How certain are you that you can solve these tasks in the next lesson?
In the next lesson in maths I can solve:

	0 Not at all certain	1	2	3	4	5 Moderately certain	6	7	8	9	10 Completely certain
all the easy tasks											
all the medium tasks											
all the hard tasks											

3: How certain are you that you can manage *not to give up* if faced with difficult problems in <algebra> in today`s lesson?
In the next lesson in maths I can manage not to give up:

	0 Not at all certain	1	2	3	4	5 Moderately certain	6	7	8	9	10 Completely certain
if I am faced with slightly difficult tasks											
if I am faced with medium difficulty tasks											
if I am faced with very difficult tasks											

Appendix L: Information letter to students / legal guardians, phase 2

To the student/legal guardian
This letter is regarding participation in the final phase of the research study "Self-efficacy beliefs". The study is part of a larger research project, "Learning regions", which is concerned with how we might facilitate teaching in schools to improve student attainment. In 2013 students in 5th and 9th grade answered questionnaires prior to the national tests in numeracy: in this final phase we wish to conduct mathematics classroom observations and short questionnaires in a selection of classes.

What
The observations will take place during 3-4 lessons in mathematics, and are planned together with the teacher. Before and after each lesson students will fill out short (maximum 5 minutes) questionnaires regarding their self-efficacy beliefs. The lessons will be conducted as normal, however with two observers present in the room. A selection of students in each class will be interviewed (maximum 10 minutes) after each lesson.
Participation in this project is voluntary, and participants may at any time withdraw from the study withouth specifying a reason. The study has obtained approval from the Data protection official for research, NSD, and the University College of Sogn & Fjordane is responsible for the project.

Why
We wish to increase our understanding regarding how to support students` learning and motivation in the mathematics classroom. The focus as such is not on individual teachers/students, but on factors that are related to student learning and motivation. We hope the knowledge gained from this study can help schools and teachers facilitate learning in school for all students.
The results from this study will be published in one or several research reports of national and international kind, and will be used to share knowledge regarding teaching and learning with teachers and in schools. All personal information gained through the study will be treated confidentially. No schools or participants will be identifiable through the research reports published.

What do I need to to?
If you consent to your child`s participation: please complete the below consent form and hand it in at school. The teacher will let you know when the observations will take place.

Kind regards,
Karin Sørlie
Principal investigator & university college lecturer, The University College of Sogn & Fjordane
DPhil student, University of Oxford, United Kingdom
Phone: 57 67 61 58
Email: karin.elisabeth.sorlie@hisf.no
Ingrid Syse
University college lecturer, The University College of Sogn & Fjordane
Phone: 57 67 61 63
Email: ingrid.syse@hisf.no

Declaration of consent:

As the legal guardian of (name of student) …….. I have received written information and I consent to my child`s participation in phase 2 of the research study ”self-efficacy beliefs”.
I consent /do not consent (cross out as appropriate) to my child being interviewed during this study.

Signature …….

Appendix M: Information letter to schools, phase 2

Dear teacher/principal:
This letter concerns participation in phase two of the study «Self-efficacy beliefs».

Thank you!
First and foremost, we wish to thank you for your participation and help so far – we greatly appreciate it! This final phase of the study involves classroom observations in mathematics. This letter contains information regarding the aim of the study, what we ask from you, and what you might gain from participating.

Aims
We wish to understand more about which factors in the mathematics classroom can support the development of students` self-efficacy beliefs. Not much is known about how this happens in the classroom, thus together we hope to create new knowledge important for researchers, school owners and not the least teachers in school. Our focus in the classroom is as such not on whether individual teachers are "good", but on what actions and interactions are related to the development of adaptive self-efficacy beliefs for students.

Observations in the classroom
We wish to observe 3-4 consecutive lessons in mathematics. We wish to be present when something new is introduced by the teacher: in other words not lessons where the focus is on repetition etc. Before each lesson we will interview the teacher (approximately 20mins). This is important to gain contextual information regarding the lesson and the students, and to know the main plan of the lesson. Prior to the first of the lessons all students will fill out a questionnaire (maximum 15 minutes) regarding their self-efficacy beliefs in mathematics (this resembles the questionnaire from 2013 but is shorter). Thereafter students will fill out short questionnaires (maximum 5 minutes) before and after each of the 3-4 lessons we observe. During the lessons two researchers will be present in the classroom. After each lesson we wish to conduct short interviews (maximum 10 minutes) with 8 students, to gain knowledge about their understanding of the lesson. We also want insight into the teacher`s understanding of lessons in hindsight: this can be discussed together with the interview prior to next lesson.

What you might gain
We greatly appreciate all the time and resources spent by schools and participants, and wish to «give something back», beyond the new knowledge we hope to create. We wish to enter into a dialogue with each school/teacher, to learn what you feel would be useful to gain from this. For instance, we could contribute feedback to the individual teacher, in regards to continued professional development. We could also contribute in a more general sense, with knowledge regarding the role of self-efficacy beliefs in education.
We will be in touch with each school directly, but appreciate any questions from you!

Kind regards,
Karin Sørlie
Principal investigator & university college lecturer, The University College of Sogn & Fjordane
DPhil student, University of Oxford, United Kingdom
Phone: 57 67 61 58
Email: karin.elisabeth.sorlie@hisf.no

Ingrid Syse
University college lecturer, The University College of Sogn & Fjordane
Phone: 57 67 61 63
Email: ingrid.syse@hisf.no

Ethics

The results from this study will be published in one or several research reports of national and international kind, and will be used to share knowledge regarding teaching and learning with teachers, school owners and other researchers.

All personal information gained through the study will be treated confidentially. No schools or participants will be identifiable through the research reports published.

Participation in this project is voluntary, and participants may at any time withdraw from the study withouth specifying a reason. The study has obtained approval from the Data protection official for research, NSD, and the University College of Sogn & Fjordane is responsible for the project.

Appendix N: Mastery Experiences Gradations of Difficulty (MEGD)

Developed by Karin Street. Words in <brackets> were changed to reflect the topic of each lesson.

For each of these questions, think about the lesson on ***<algebra>*** *which has just ended.*

1: How true is it that you succeeded in learning what the teacher taught you about <algebra> in today`s lesson?

In today`s lesson I succeeded in learning:

	0 Not at all true	1	2	3	4	5 Moderately true	6	7	8	9	10 Completely true
at least a little bit of what the teacher taught us											
at least half of what the teacher taught us											
everything the teacher taught us											

2: During the last lesson, some <algebra> tasks might have seemed easy and some might have seemed hard *to you*. How true is it that you succeeded in solving these tasks?

In today`s lesson I succeeded in solving:

	0 Not at all true	1	2	3	4	5 Moderately true	6	7	8	9	10 Completely true
all the tasks that seemed easy to me											
all the tasks that seemed of medium difficulty to me											
all the tasks that seemed hard to me											

3: How true is it that you managed *not to give up* when faced with difficult problems in <algebra> in today`s lesson?

In today`s lesson I managed not to give up:

	0 Not at all true	1	2	3	4	5 Moderately true	6	7	8	9	10 Completely true
when I was faced with slightly difficult tasks											
when I was faced with medium difficulty tasks											
when I was faced with very difficult tasks											

Appendix O:

Self-Efficacy Task instrument

Developed by Karin Street. The words in <brackets> were changed to reflect the main task of each lesson.

1) One of the tasks you worked on today was <task description>. How certain do you now feel you can solve similar tasks? *(Tick one box)*										
0 Not at all certain	1	2	3	4	5 Moderately certain	6	7	8	9	10 Completely certain

Appendix P: Ability Conceptions Questionnaire

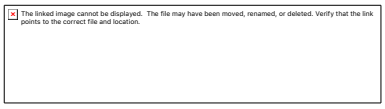
Developed by Carol Dweck (2008). Adapted by Karin Street.

What does it mean to be good at maths? There are no right or wrong answers here, we are interested in your opinions. With each sentence indicate whether you agree or disagree by choosing a number between 1 and 6.

How much do you agree that:

	1 Strongly agree	2 Agree	3 Somewhat agree	4 Somewhat disagree	5 Disagree	6 Strongly disagree
People CAN`T really do much to change if they are good at maths or not.						
Even how good they really are at maths, people can change substantially						
No matter who someone is, they can always become better at maths if they try						
To be honest, people CAN`T really change how good they are at maths						
No matter who someone is, they can change how smart they are a lot						
How smart someone is, is something they CAN`T change very much						
People can always substantially change how smart they are						
People can learn new things, but they CAN`T really change how smart they are						

Appendix Q: Norwegian instruments, phase 1



Spørreskjema elever trinn 8

På dette spørreskjemaet er det viktig at du leser spørsmålene nøye, og at du svarer ærlig. Det er ingen andre enn forskerne som kommer til å se svarene dine, og det er ingen som får vite hvem du er.

Spør læreren din om hjelp om det er noe du er usikker på.
Lykke til!

1: Bakgrunns-spørsmål. Her følger noen spørsmål om deg. Det er bare forskerne som får se dem, og ingen andre får vite hvem du er.						
	Skole					
	Klasse					
	Navn					
	Alder	År		Måneder		(eksempel År: 14 Måneder: 8)
	Gutt eller jente (sett kryss)	Gutt		Jente		

Øve spørsmål:

Gjør dette spørsmålet sammen med læreren din.
Tenk deg at du blir bedt om å løfte en sekk akkurat nå. Hvor sikker er du på at du kan løfte sekken?

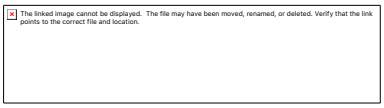
For hver sekk, kryss av for tallet som beskriver hvor sikker du er. Tallene er mellom 0 og 10, der (0) = "det er jeg absolutt ikke sikker på om jeg kan løfte", (5) = "det er jeg middels sikker på at jeg kan løfte", og (10) = "det er jeg fullstendig sikker på at jeg kan løfte".

En som for eksempel er litt under middels sikker på at han kan løfte en sekk på 50kg, kan krysse av for tallet 4. Prøv selv under.

Jeg kan klare å løfte, akkurat nå:

	0 Absolutt ikke sikker	1	2	3	4	5 Middels sikker	6	7	8	9	10 Fullstendig sikker
En sekk som veier 1 kg											
En sekk som veier 10 kg											
En sekk som veier 50 kg											
En sekk som veier 100 kg											

Spørreskjema elever trinn 8



Nå følger forskjellige setninger som handler om de nasjonale prøvene i regning i år. Du skal krysse av for tallet som best beskriver hvor sikker du er på at du kan gjøre det som står. Tallene er mellom 0 og 10, der (0) betyr «det er jeg absolutt ikke sikker på om jeg kan», (5) betyr «det er jeg middels sikker på at jeg kan» og (10) betyr «det er jeg helt sikker på at jeg kan».

2: Det er over 50 oppgaver i de nasjonale prøvene i regning. Hvor sikker er du på at du vil kunne løse minst et visst antall av oppgavene?

På de nasjonale prøvene i år vil jeg klare å løse:

	0 Absolutt ikke sikker	1	2	3	4	5 Middels sikker	6	7	8	9	10 Fullstendig sikker
minst 5 av oppgavene											
minst 10 av oppgavene											
minst 25 av oppgavene											
minst 40 av oppgavene											
alle oppgavene											

3: På de nasjonale prøvene kan man oppnå forskjellige karakterer, avhengig av hvor bra man gjør det på prøven. Man kan oppnå mellom 1 (laveste karakter) og 5 (høyeste karakter). Hvor sikker er du på at du vil kunne oppnå minst hver karakter?

På de nasjonale prøvene i år vil jeg klare å oppnå:

	0 Absolutt ikke sikker	1	2	3	4	5 Middels sikker	6	7	8	9	10 Fullstendig sikker
minst karakter 1											
minst karakter 2											
minst karakter 3											
minst karakter 4											
karakter 5											

Spørreskjema elever trinn 8

4: På de nasjonale prøvene kan noen oppgaver oppleves som lette, og noen som vanskeligere. Hvor sikker er du på at du vil kunne løse disse oppgavene?

På de nasjonale prøvene i år vil jeg klare å løse:

	0 Absolutt ikke sikker	1	2	3	4	5 Middels sikker	6	7	8	9	10 Fullstendig sikker
alle de lette oppgavene											
alle de middels vanskelige oppgavene											
alle de vanskelige oppgavene											

5: Noen synes det kan være vanskelig å konsentrere seg når man har prøver i regning. Hvor sikker er du på at du vil kunne konsentrere deg i løpet av de nasjonale prøvene i regning i år?

På de nasjonale prøvene i år vil jeg klare å konsentrere meg:

	0 Absolutt ikke sikker	1	2	3	4	5 Middels sikker	6	7	8	9	10 Fullstendig sikker
i hvert fall en liten stund											
i hvert fall halve prøven											
hele prøven											

6: Det kan være lett å gi opp om man møter vanskelige spørsmål på en matteprøve. Hvor sikker er du på at du vil klare å *ikke gi opp* nar du møter vanskelige oppgaver på de nasjonale prøvene i regning i år?

På de nasjonale prøvene i år vil jeg klare å la være å gi opp:

	0 Absolutt ikke sikker	1	2	3	4	5 Middels sikker	6	7	8	9	10 Fullstendig sikker
når jeg møter litt vanskelige oppgaver											
når jeg møter ganske vanskelige oppgaver											
når jeg møter veldig vanskelige oppgaver											

Appendix R: Norwegian instruments, phase 2

(Pre-observation questionnaire, followed by student lesson questionnaire)

HØGSKULEN i SOGN OG FJORDANE

Navn:

Det er viktig at du leser spørsmålene nøye og svarer ærlig. Det er bare forskerne som får se svarene dine, og ingen andre får vite hvem du er.

Spør om hjelp om det er noe du er usikker på!

Øve-spørsmål: *Gjør øve-spørsmålet sammen med læreren*

Tenk deg at du blir bedt om å løfte en sekk akkurat nå. Hvor sikker er du på at du kan løfte sekken?

For hver linje, kryss av for tallet som beskriver hvor sikker du er. Tallene er mellom 0 og 10, der (0) = "det er jeg absolutt ikke sikker på om jeg kan løfte", (5) = "det er jeg middels sikker på at jeg kan løfte", og (10) = "det er jeg fullstendig sikker på at jeg kan løfte".

En som for eksempel er litt under middels sikker på at han kan løfte en sekk på 50kg, kan krysse av for tallet 4. Prøv selv under.

Jeg kan klare å løfte, akkurat nå:

	0 Absolutt ikke sikker	1	2	3	4	5 Middels sikker	6	7	8	9	10 Fullstendig sikker
En sekk som veier 1 kg											
En sekk som veier 10 kg											
En sekk som veier 50 kg											
En sekk som veier 100 kg											

Resten av spørsmålene skal du gjøre alene. Sett 1 kryss for hver linje.

Disse spørsmålene handler om de neste timene i matematikk. Velg det tallet som best beskriver hvor sikker du er.

1: I de neste matematikktimene skal dere lære om et nytt emne. Hvor sikker er du på at du kan lære dette med en viss mengde hjelp fra læreren?

I løpet av de neste mattetimene kan jeg lære et nytt emne:

	0 Absolutt ikke sikker	1	2	3	4	5 Middels sikker	6	7	8	9	10 Helt sikker
om jeg får mye hjelp fra læreren											
om jeg får litt hjelp fra læreren											
uten hjelp fra læreren i det hele tatt											

2: Når man lærer nye emner i matematikk jobber man ofte med å løse oppgaver. Hvor sikker er du på at du kan klare å løse et visst antall?

I hver av de neste mattetimene kan jeg løse:

	0 Absolutt ikke sikker	1	2	3	4	5 Middels sikker	6	7	8	9	10 Helt sikker
i hvert fall noen få av oppgavene jeg prøver											
i hvert fall halvparten av oppgavene jeg prøver											
alle oppgavene jeg prøver											

3: Når man løser nye oppgaver i matematikk kan noen oppgaver virke lette og noen vanskelige *for deg*. Hvor sikker er du på at du kan løse disse oppgavene?

I de neste mattetimene kan jeg løse:

	0 Absolutt ikke sikker	1	2	3	4	5 Middels sikker	6	7	8	9	10 Helt sikker
alle de lette oppgavene											
alle de middels vanskelige oppgavene											
alle de vanskelige oppgavene											

4: Av og til kan det være vanskelig å konsentrere seg når man jobber med matematikk. Hvor sikker er du på at du kan klare å konsentrere deg når du jobber med å løse oppgaver i de neste mattetimene?

Jeg kan klare å konsentrere meg:

	0 Absolutt ikke sikker	1	2	3	4	5 Middels sikker	6	7	8	9	10 Helt sikker
i hvert fall en liten stund											
i hvert fall halvparten av tiden											
hele tiden											

5: Det kan være lett å gi opp når man møter vanskelige nye oppgaver. Hvor sikker er du på at du kan klare å *ikke gi opp* om du møter vanskelige oppgaver i algebra i de neste mattetimene?
I de neste matematikktimene kan jeg klare å *ikke gi opp*:

	0 Absolutt ikke sikker	1	2	3	4	5 Middels sikker	6	7	8	9	10 Helt sikker
om jeg møter litt vanskelige oppgaver											
om jeg møter middels vanskelige oppgaver											
om jeg møter veldig vanskelige oppgaver											

6: Hva betyr det å være god i matematikk? Det finnes ingen riktige eller gale svar her, vi vil vite hva du synes. For hver setning, skriv om du er enig eller uenig ved å sette ett kryss på skalaen som går fra 1-6.

Hvor enig er du i at:

	1 Veldig enig	2 Enig	3 Litt enig	4 Litt uenig	5 Uenig	6 Veldig uenig
Folk kan IKKE gjøre mye for å endre om de er gode i matte eller ikke						
Selv hvor god man egentlig er i matte kan man endre mye						
Uansett hvem noen er, kan de alltid bli bedre i matte om de prøver						
For å være ærlig, så kan man egentlig IKKE endre hvor god man er i matte						
Uansett hvem noen er, så kan de endre hvor smarte de er mye						
Hvor smart noen er, er noe de IKKE kan endre veldig mye						
Folk kan alltid endre hvor smarte de er mye						
Folk kan lære nye ting, men de kan egentlig IKKE endre hvor smarte de er						

HØGSKULEN i SOGN OG FJORDANE

Navn:

(This page is intentionally left blank, as it was in the original instrument)

For hvert av disse spørsmålene, tenk på mattetimen om geometri som skal til å starte.

1: I mattetimen i dag skal du lære om geometri. Hvor sikker er du på at du kan lære dette med en viss mengde hjelp fra læreren?

I den neste mattetimen kan jeg lære geometri:

	0 Absolutt ikke sikker	1	2	3	4	5 Middels sikker	6	7	8	9	10 Helt sikker
om jeg får mye hjelp fra læreren											
om jeg får litt hjelp fra læreren											
uten hjelp fra læreren i det hele tatt											

2: Når du løser matteoppgaver i geometri kan noen oppgaver virke lette og noen vanskelige *for deg.* Hvor sikker er du på at du kan løse disse oppgavene i den neste mattetimen?
I den neste mattetimen kan jeg løse:

	0 Absolutt ikke sikker	1	2	3	4	5 Middels sikker	6	7	8	9	10 Helt sikker
alle de lette oppgavene											
alle de middels vanskelige oppgavene											
alle de vanskelige oppgavene											

3: Hvor sikker er du på at du kan klare å *ikke gi opp* om du møter vanskelige oppgaver i geometri i den neste mattetimen?
I den neste mattetimen kan jeg klare å *ikke gi opp*:

	0 Absolutt ikke sikker	1	2	3	4	5 Middels sikker	6	7	8	9	10 Helt sikker
om jeg møter litt vanskelige oppgaver											
om jeg møter middels vanskelige oppgaver											
om jeg møter veldig vanskelige oppgaver											

For hvert av disse spørsmålene, tenk på mattetimen om geometri som akkurat er ferdig.

4: Hvor sant er det at du klarte å lære det læreren underviste om geometri i mattetimen i dag?
I mattetimen i dag klarte jeg å lære:

	0 Absolutt ikke sant	1	2	3	4	5 Middels sant	6	7	8	9	10 Helt sant
i hvert fall litt av det læreren underviste											
i hvert fall halvparten av det læreren underviste											
alt læreren underviste											

5: I mattetimen i dag kan noen oppgaver i geometri ha virket lette og noen ha virket vanskelige *for deg*. Hvor sant er det at du klarte å løse disse oppgavene?
I mattetimen i dag klarte jeg å løse:

	0 Absolutt ikke sant	1	2	3	4	5 Middels sant	6	7	8	9	10 Helt sant
alle oppgavene som virket lette for meg											
alle oppgavene som virket middels vanskelige for meg											
alle oppgavene som virket vanskelige for meg											

6: Hvor sant er det at du klarte å *ikke gi opp* om du møtte vanskelige oppgaver i geometri i mattetimen i dag?
I mattetimen i dag klarte jeg å *ikke gi opp*:

	0 Absolutt ikke sant	1	2	3	4	5 Middels sant	6	7	8	9	10 Helt sant
om jeg møtte litt vanskelige oppgaver											
om jeg møtte middels vanskelige oppgaver											
om jeg møtte veldig vanskelige oppgaver											

7) En av oppgavene du jobbet med i dag var å forklare geometriske figurer. Hvor sikker er du på at du nå kan løse lignende oppgaver? *(Sett 1 kryss)*

0 Absolutt ikke sikker	1	2	3	4	5 Middels sikker	6	7	8	9	10 Helt sikker