

**From teacher-regulation to self-regulation in early childhood:
An analysis of Tools of the Mind's curricular effects**

Alex Baron

Dissertation for the Doctor of Philosophy

Hilary Term 2017

Supervised by Dr. Malmberg and Dr. Evangelou



Dedication

Let's just call him Anthony. On my first day of preschool teaching, I witnessed the shimmering smile of a child who could not control himself but excelled in controlling me. My teaching day was largely dictated by the impulses surging through his four-foot frame. As a teacher, I aimed to help children like Anthony master foundational literacy, numeracy, and the intangible but crucial skills involved in "school readiness." During my two years of preschool teaching, I skillfully executed the first two and failed miserably at the third. This DPhil represents my redemptive attempt to address aim number three so that I, and all teachers, can do better for our kids.

Acknowledgements

The list of people who have supported me through this process would probably be longer than the dissertation itself. Given spatial and temporal constraints, I would like to especially thank the subset of wonderful humans described below. To those not explicitly mentioned in the acknowledgements section, I offer my deep gratitude for whatever time and energy you contributed to this work. Much love.

First and foremost, I would like to thank my supervisors, Dr. Lars-Erik Malmberg and Dr. Maria Evangelou, who provided lots of “sisu” and “ελαστικότητα” during this project. Without their unwavering support and clear guidance, I would likely still be struggling to successfully execute a t-test on SPSS. I am boundlessly grateful to them both.

Secondly, I would like to thank my family for helping me self-regulate throughout the last several weeks, months, and years. In addition to my research supervisors, I also had my parents, sister, and extended family to serve as my emotional supervisors during the occasionally treacherous doctoral marathon. One simply could not ask for better kinfolk.

Thirdly, I would like to thank my Oxford friends, whose physical presence came and went during the four years but whose emotional presence never budged. In particular, I’d like to thank Shea, Kerry, Moizza, Pete, Sheiry, Aneil, JT, Nina, Benedict, and GJ for being stunningly reliable sources of challenge, support, and general tomfoolery.

Fourthly, I would like to thank both the Peabody Research Institute research team, who generously provided both guidance and data throughout this project, as well as the European Social and Economic Research Council and Marshall Scholarship, who generously provided funding throughout this project.

Finally, I would like to thank my incredible students and fellow educators who inspired this paper as much as they inspired me. I’m sorry if this thesis did not crack the self-regulation code specifically or the effective education code more generally, but I look forward to a lifetime of figuring it out alongside you all.

Abstract

The aim of my DPhil is to identify educational practices predictive of students' self-regulation development during early childhood. Specifically, I will analyze the Tools of the Mind preschool curriculum (Tools), which emphasizes students' self-regulation cultivation as its paramount aim.

Since its development in 1993, Tools has spread to schools in the United States, Canada, and South America. In the face of Tools' proliferation, two questions emerge: does Tools significantly improve children's self-regulation skills? And, if so, then which of its effective elements could be applied across various educational contexts?

This dissertation contains two studies. In the first, I will systematically review extant Tools research and then execute a multilevel meta-analysis of the quantitative results. Study one serves three purposes: 1) to identify all studies in the existing Tools evidence base, 2) to estimate an aggregate curricular effect, and 3) to determine how that effect varies across contexts and student characteristics. Thus, study one will assess whether Tools, at the curricular level, improves students' self-regulation.

By contrast, study two will involve more granular analyses of the discrete learning activities that collectively comprise Tools. Specifically, study two will analyze child-level self-regulation and teacher-level Tools implementation data for 1145 preschool children in 80 classrooms across six American school districts. I will employ multilevel structural equation models to assess which Tools activities are associated with students' self-regulation growth, which are associated with decline, and which exhibit no association at all.

Ultimately, this dissertation features the first Tools meta-analysis as well as the first analysis of specific Tools instructional activities. It is hoped that these analyses will identify educational practices predictive of self-regulation development both within and beyond the Tools curricular context.

Table of contents

CHAPTER 1: DISSERTATION OVERVIEW.....	18
1.1 WHAT IS SELF-REGULATION?	18
1.2 RESEARCH RATIONALE	21
1.3 RESEARCH FOCUS: TOOLS OF THE MIND.....	24
1.4 POLICY CONTEXT: WHY IS IT IMPORTANT TO EVALUATE TOOLS?.....	26
1.5 THEORETICAL FRAMEWORKS FOR THIS DISSERTATION	27
1.6 STRUCTURE OF THE ANALYSES IN THIS DISSERTATION	27
1.7 OVERALL DISSERTATION STRUCTURE.....	29
 CHAPTER 2: VYGOTSKY, THE CULTURAL-HISTORICAL THEORY, AND TOOLS OF THE MIND	33
2.1 VYGOTSKY’S RESEARCH CONTRIBUTION.....	33
2.2 CULTURAL-HISTORICAL THEORY	33
2.3 VYGOTSKY’S INFLUENCE ON TOOLS OF THE MIND	39
2.4 TOOLS’ THEORY OF CHANGE	40
 CHAPTER 3: BRONFENBRENNER’S BIOECOLOGICAL MODEL	46
3.1 OTHER THEORETICAL ACCOUNTS OF SELF-REGULATION DEVELOPMENT	46
3.2 BRONFENBRENNER’S ECOLOGICAL MODEL.....	48
 CHAPTER 4: WHY IS SELF-REGULATION IMPORTANT?	56

4.1	SELF-REGULATION AND ACADEMIC ACHIEVEMENT.....	56
4.2	COMPOUNDING BENEFITS OF SELF-REGULATION THROUGHOUT THE LIFE COURSE.....	61
CHAPTER 5: BIOLOGICAL AND CONTEXTUAL FACTORS THAT INFLUENCE SELF-REGULATION DEVELOPMENT		63
5.1	GENES AND NEUROBIOLOGY OF SELF-REGULATION	63
5.2	CONTEXTUAL FACTORS THAT AFFECT SELF-REGULATION.....	66
5.3	CHAPTER SUMMARY	82
CHAPTER 6: SELF-REGULATION INTERVENTIONS		84
6.1	INCREDIBLE YEARS TEACHER CLASSROOM MANAGEMENT PROGRAM.....	84
6.2	CHICAGO SCHOOL READINESS PROJECT	88
6.3	RESEARCH-BASED, DEVELOPMENTALLY INFORMED (REDI) PROGRAM.....	89
6.4	PROMOTING ALTERNATIVE THINKING STRATEGIES (PATHS) PROGRAM.....	90
6.5	SUMMARY OF SUPPLEMENTARY PROGRAMS ASSOCIATIONS WITH CHILDREN’S SELF-REGULATION	92
CHAPTER 7: BACKGROUND ON EARLY CHILDHOOD CURRICULA		94
7.1	HIGHSCOPE EARLY CHILDHOOD CURRICULUM	94
7.2	CREATIVE CURRICULUM.....	96
7.3	BRIGHT BEGINNINGS.....	98
7.4	MONTESORI	99
7.5	TOOLS OF THE MIND.....	101

CHAPTER 8: METHODOLOGY FOR STUDY ONE.....	105
8.1 RATIONALE FOR THIS RESEARCH SYNTHESIS PROJECT	105
8.2 RESEARCH SYNTHESIS: SYSTEMATIC REVIEWS AND META-ANALYSES	106
8.3 DRAWBACKS OF RESEARCH SYNTHESIS METHODS	113
8.4 THE APPROACH OF THE PRESENT REVIEW.....	115
CHAPTER 9: RESULTS FOR STUDY ONE.....	135
9.1 RESULTS OF THE SYSTEMATIC SEARCH.....	135
9.2 CHARACTERISTICS OF EXCLUDED STUDIES	137
9.3 ONGOING STUDIES	138
9.4 NARRATIVE SUMMARY OF EACH INCLUDED STUDY	139
9.5 OVERVIEW OF THE CHARACTERISTICS OF INCLUDED STUDIES.....	145
9.6 RISK OF BIAS IN INCLUDED STUDIES.....	147
9.7 MULTILEVEL META-ANALYTIC FINDINGS ON THE INTERVENTION'S EFFECT.....	152
9.8 HETEROGENEITY IN THE COMPOSITE EFFECT SIZES.....	159
9.9 MODERATION ANALYSIS.....	160
9.10 SENSITIVITY ANALYSIS.....	161
9.11 ROBUSTNESS CHECK WITH ROBUST VARIANCE ESTIMATION	163
9.12 CHAPTER SUMMARY	164
CHAPTER 10: DISCUSSION FOR STUDY ONE.....	165
10.1 SUMMARY OF THE MAIN RESULTS.....	165

10.2	QUALITY OF THE EVIDENCE BASE	169
10.3	DIFFERENCES BETWEEN PROTOCOL AND REVIEW.....	170
10.4	STRENGTHS.....	171
10.5	LIMITATIONS.....	172
10.6	SUGGESTIONS FOR FUTURE RESEARCH	176
10.7	CHAPTER SUMMARY	179
CHAPTER 11: METHODOLOGY FOR STUDY TWO.....		181
11.1	DATA SAMPLE	181
11.2	DESIGN AND PROCEDURES OF THE PRI STUDY.....	181
11.3	ANALYTIC APPROACH: SECONDARY DATA ANALYSIS.....	186
11.4	ETHICS	187
11.5	SELF-REGULATION MEASURES.....	187
11.6	FACTOR ANALYSIS.....	199
11.7	TOOLS ACTIVITY IMPLEMENTATION DATA.....	208
11.8	COVARIATES.....	210
11.9	MISSING DATA TREATMENT	212
11.10	RESEARCH QUESTIONS AND HYPOTHESES FOR STUDY TWO.....	214
11.11	STATISTICAL APPROACH: MULTILEVEL STRUCTURAL EQUATION MODELING	218
11.12	METHODOLOGICAL SUMMARY	225
CHAPTER 12: RESULTS FOR STUDY TWO.....		227

12.1	RESEARCH QUESTION ONE: CHILDREN’S SELF-REGULATORY DEVELOPMENT BETWEEN PRE-KINDERGARTEN AND FIRST GRADE	227
12.2	RESEARCH QUESTION TWO: ‘TOOLS’ EFFECT ON SELF-REGULATION VIS-À-VIS COMPARATOR CURRICULA.....	239
12.3	RESEARCH QUESTION THREE: SUBJECT-SPECIFIC TOOLS ACTIVITY BLOCKS AND SELF-REGULATION	247
12.4	RESEARCH QUESTION FOUR: SPECIFIC TOOLS ACTIVITIES AND SELF-REGULATION	258
12.5	SUMMARY OF RESULTS FOR STUDY TWO	282
CHAPTER 13: DISCUSSION FOR STUDY TWO		284
13.1	RESEARCH QUESTION ONE: CHILDREN’S SELF-REGULATORY DEVELOPMENT BETWEEN PRE-KINDERGARTEN AND FIRST GRADE	284
13.2	RESEARCH QUESTION TWO: ‘TOOLS’ EFFECT ON SELF-REGULATION VIS-À-VIS COMPARATOR CURRICULA.....	288
13.3	RESEARCH QUESTION THREE: SUBJECT-SPECIFIC TOOLS ACTIVITY BLOCKS AND SELF-REGULATION	289
13.4	RESEARCH QUESTION FOUR: SPECIFIC TOOLS ACTIVITIES AND SELF-REGULATION DEVELOPMENT	298
13.5	BRIEF SUMMARY OF THE STUDY TWO FINDINGS	299
13.6	STRENGTHS.....	300
13.7	LIMITATIONS.....	301
13.8	DIRECTIONS FOR FUTURE RESEARCH	308
13.9	CHAPTER SUMMARY	311

CHAPTER 14: CONCLUSION	313
14.1 BRIEF SUMMARY OF STUDY ONE AND TWO FINDINGS	313
14.2 CONTRIBUTIONS TO THE LITERATURE BASE	314
14.3 IMPLICATIONS FOR POLICY AND PRACTICE	315
REFERENCES.....	316
APPENDICES	357

List of tables

TABLE 1: INCLUDED STUDIES AND THE RECORDS ASSOCIATED WITH THOSE STUDIES	136
TABLE 2: EXCLUDED STUDIES AND REASONS FOR THEIR EXCLUSION	138
TABLE 3: CHARACTERISTICS OF ONGOING STUDIES.....	138
TABLE 4: MULTILEVEL META-ANALYSIS RESULTS ACROSS THE FOUR OUTCOME MEASURES.....	153
TABLE 5: HETEROGENEITY ANALYSIS FOR THE COMPOSITE EFFECT SIZES	160
TABLE 6: SENSITIVITY ANALYSIS TO REMOVE STUDIES WITH A HIGH-RISK OF BIAS.....	162
TABLE 7: SENSITIVITY ANALYSIS FOR TOOLS AS A COMBINED INTERVENTION.....	163
TABLE 8: ROBUSTNESS CHECK WITH ROBUST VARIANCE ESTIMATION	164
TABLE 9: COHORT 1 AND 2 DIFFERENCES BETWEEN TOOLS AND CONTROL CLASSROOMS	182
TABLE 10: CORSI BLOCKS DESCRIPTIVE STATISTICS AT EACH TIME POINT.....	189
TABLE 11: DIMENSIONAL CHANGE CARD SORT DESCRIPTIVE STATISTICS ACROSS TIME POINTS	190
TABLE 12: PEG TAPPING DESCRIPTIVE STATISTICS ACROSS TIME POINTS	191
TABLE 13: HEADS-TOES-KNEES-SHOULDERS DESCRIPTIVE STATISTICS ACROSS TIME POINTS.....	192
TABLE 14: COPY DESIGN DESCRIPTIVE STATISTICS ACROSS TIME POINTS	193
TABLE 15: RESEARCHER-REPORTED SAR DESCRIPTIVE STATISTICS ACROSS TIME POINTS.....	195
TABLE 16: TEACHER-REPORTED CFBRs DESCRIPTIVE STATISTICS ACROSS TIME POINTS	197
TABLE 17: CFA RESULTS FOR THE LATENT EXECUTIVE FUNCTION CONSTRUCT ACROSS TIME POINTS.....	204
TABLE 18: CFA RESULTS FOR THE LATENT RESEARCHER-REPORTED SAR CONSTRUCT ACROSS TIME POINTS	205
TABLE 19: CFA RESULTS FOR THE LATENT TEACHER-REPORTED CFBRs CONSTRUCT ACROSS TIME POINTS.....	205

TABLE 20: CONFIGURAL, WEAK, AND STRONG MEASUREMENT INVARIANCE RESULTS FOR THE EXECUTIVE FUNCTION CONSTRUCT.....	207
TABLE 21: CONFIGURAL, WEAK, AND STRONG MEASUREMENT INVARIANCE RESULTS FOR THE SAR CONSTRUCT	207
TABLE 22: CONFIGURAL, WEAK, AND STRONG MEASUREMENT INVARIANCE RESULTS FOR THE CFBRs CONSTRUCT	207
TABLE 23: INTRA-CLASS CORRELATION (CHILDREN IN CLASSROOMS) VALUES FOR EACH OF THE OBJECTIVE SELF- REGULATION MEASURES.....	221
TABLE 24: UNSTANDARDIZED INTERCEPT AND GROWTH RATE ESTIMATES FOR THE CORSI BLOCKS TASK.....	228
TABLE 25: UNSTANDARDIZED INITIAL STATUS AND GROWTH RATE ESTIMATES FOR THE HTKS TASK	230
TABLE 26: UNSTANDARDIZED INITIAL STATUS AND GROWTH RATE ESTIMATES FOR THE DCCS TASK.....	231
TABLE 27: UNSTANDARDIZED INITIAL STATUS AND GROWTH RATE ESTIMATES FOR THE COPY DESIGN TASK.....	233
TABLE 28: UNSTANDARDIZED INITIAL STATUS AND GROWTH RATE ESTIMATES FOR THE PEG TAPPING TASK.....	234
TABLE 29: UNSTANDARDIZED GROWTH RATE FOR THE SAR CONSTRUCT	236
TABLE 30: UNSTANDARDIZED GROWTH RATE ESTIMATES FOR THE CFBRs CONSTRUCT	237
TABLE 31: INTERACTION EFFECTS FOR CHILD GENDER, ELL, AND SEN STATUS	245
TABLE 32: COHORT SENSITIVITY ANALYSIS RESULTS ACROSS ALL THREE OUTCOME MEASURES	246
TABLE 33: TEACHER-REPORTED IMPLEMENTATION FREQUENCY DESCRIPTIVE STATISTICS FOR THE TOOLS ACTIVITY BLOCKS	249
TABLE 34: BETA COEFFICIENTS FOR THE MAKE-BELIEVE PLAY (MBP) BLOCK ON ALL THREE OUTCOME MEASURES	251
TABLE 35: BETA COEFFICIENTS FOR THE LITERACY ACTIVITY BLOCK ON ALL THREE OUTCOME MEASURES.....	252
TABLE 36: BETA COEFFICIENTS FOR THE MATH ACTIVITY BLOCK ON ALL THREE OUTCOME MEASURES	253
TABLE 37: BETA COEFFICIENTS FOR THE SCIENCE BLOCK ON ALL THREE OUTCOME MEASURES.....	254

TABLE 38: BETA COEFFICIENTS FOR THE ATTENTION-FOCUSING ACTIVITY BLOCK ON ALL THREE OUTCOME MEASURES	254
TABLE 39: BETA COEFFICIENTS FOR THE INTRODUCTION ACTIVITY BLOCK ON ALL THREE OUTCOME MEASURES.....	255
TABLE 40: SENSITIVITY ANALYSIS WITH RESEARCHER-OBSERVED FIDELITY DATA	256
TABLE 41: RAW CORRELATIONS AMONG EACH TOOLS ACTIVITY AND THE THREE OUTCOME MEASURES.....	258
TABLE 42: BETA COEFFICIENTS FOR MAKE-BELIEVE PLAY PLANNING (ACT) ON THE THREE OUTCOME MEASURES ..	261
TABLE 43: BETA COEFFICIENTS FOR MAKE-BELIEVE PLAY PRACTICE ON THE THREE OUTCOME MEASURES.....	262
TABLE 44: BETA COEFFICIENTS FOR THE MAKE-BELIEVE PLAY CENTERS ON THE THREE OUTCOME MEASURES.....	262
TABLE 45: BETA-COEFFICIENTS FOR FREEZE GAME ON THE THREE OUTCOME MEASURES	264
TABLE 46: BETA COEFFICIENTS FOR FREEZE ON NUMBER ON THE THREE OUTCOME MEASURES	265
TABLE 47: BETA COEFFICIENTS FOR PARTNER FREEZE ON THE THREE OUTCOME MEASURES	265
TABLE 48: BETA COEFFICIENTS FOR TWO-STEP FREEZE ON THE THREE OUTCOME MEASURES	266
TABLE 49: BETA COEFFICIENTS FOR THE PATTERN MOVEMENT GAME ON THE THREE OUTCOME MEASURES.....	266
TABLE 50: BETA COEFFICIENTS FOR NUMBER FOLLOW THE LEADER ON THE THREE OUTCOME MEASURES	267
TABLE 51: BETA COEFFICIENTS FOR MOUSE TRAP ON THE THREE OUTCOME MEASURES.....	267
TABLE 52: BETA COEFFICIENTS FOR THE WOLF ACTIVITY ON THE THREE OUTCOME MEASURES.....	268
TABLE 53: BETA COEFFICIENTS FOR PRETEND TRANSITION ON THE THREE OUTCOME MEASURES	269
TABLE 54: BETA COEFFICIENTS FOR THE ATTENTION FOCUS ACTIVITY ON THE THREE OUTCOME MEASURES	270
TABLE 55: BETA COEFFICIENTS FOR THE PUZZLES, MANIPULATIVES, AND BLOCKS ACTIVITY ON THE THREE OUTCOME MEASURES.....	271
TABLE 56: BETA COEFFICIENTS FOR REMEMBER AND REPLICATE ON THE THREE OUTCOME MEASURES	271

TABLE 57: BETA COEFFICIENTS FOR I HAVE WHO HAS COLORS, NUMBERS, AND SHAPES ON THE THREE OUTCOME MEASURES.....	272
TABLE 58: BETA COEFFICIENTS FOR MAKING COLLECTIONS ON THE THREE OUTCOME MEASURES	273
TABLE 59: BETA COEFFICIENTS FOR MATH MEMORY ON THE THREE OUTCOME MEASURES	273
TABLE 60: BETA COEFFICIENTS FOR NUMBER LINE HOPSCOTCH ON THE THREE OUTCOME MEASURES.....	274
TABLE 61: BETA COEFFICIENTS FOR NUMERALS GAME ON THE THREE OUTCOME MEASURES.....	274
TABLE 62: BETA COEFFICIENTS FOR VENGER DRAWING ON THE THREE OUTCOME MEASURES.....	275
TABLE 63: BETA COEFFICIENTS FOR ATTRIBUTE GAME ON THE THREE OUTCOME MEASURES.....	276
TABLE 64: BETA COEFFICIENTS FOR PATTERNS WITH MANIPULATIVES ON THE THREE OUTCOME MEASURES	276
TABLE 65: SENSITIVITY ANALYSIS WITH THE COMPOSITE TOOLS TEACHER FIDELITY VARIABLE ON THE THREE OUTCOME MEASURES	277
TABLE 66: OBSERVED EFFECTS FOR THE PLAY, MATH, AND ATTENTION-FOCUSING ACTIVITIES ACROSS THE THREE OUTCOME MEASURES	279
TABLE 67: COVARIATE EFFECTS ON THE EXECUTIVE FUNCTION INTERCEPT AND GROWTH PARAMETER ESTIMATES	280
TABLE 68: COVARIATE EFFECTS FOR THE CFBRS INTERCEPT AND GROWTH PARAMETER ESTIMATES	281
TABLE 69: COVARIATE EFFECTS FOR THE SAR INTERCEPT AND GROWTH PARAMETER ESTIMATES	281

List of figures

FIGURE 1: TOOLS THEORY OF CHANGE	38
FIGURE 2: TOOLS OF THE MIND PLAY PLAN	40
FIGURE 3: TOOLS OF THE MIND SCAFFOLDED WRITING EXAMPLE.....	41
FIGURE 4: BANDURA'S RECIPROCAL DETERMINISM MODEL.....	45
FIGURE 5: BRONFENBRENNER'S ECOLOGICAL MODEL.....	47
FIGURE 6: ADAPTED BRONFENBRENNER MODEL TO BE USED FOR THIS DISSERTATION.....	50
FIGURE 7: DATA STRUCTURE IN MULTILEVEL META-ANALYSIS.....	123
FIGURE 8: SAMPLE FUNNEL PLOT WITH EVIDENCE OF PUBLICATION BIAS.....	130
FIGURE 9: SAMPLE FUNNEL PLOT WITHOUT EVIDENCE OF PUBLICATION BIAS.....	131
FIGURE 10: SYSTEMATIC REVIEW FLOWCHART	135
FIGURE 11: RISK OF BIAS SUMMARY ACROSS ALL INCLUDED STUDIES.....	150
FIGURE 12: FOREST PLOT FOR ASSESSOR-REPORTED SELF-REGULATION EFFECT SIZES ACROSS STUDIES.....	153
FIGURE 13: FOREST PLOT FOR TASK-BASED SELF-REGULATION EFFECT SIZES ACROSS STUDIES	154
FIGURE 14: FOREST PLOT FOR LITERACY EFFECT SIZES ACROSS STUDIES	156
FIGURE 15: FOREST PLOT FOR THE MATH EFFECT SIZES ACROSS STUDIES.....	157
FIGURE 16: CHRONOLOGICAL PROGRESSION OF THE PEABODY RESEARCH INSTITUTE'S (PRI) TOOLS STUDY.....	182
FIGURE 17: UNCONDITIONAL GROWTH MODEL WITH A LATENT INTERCEPT, SLOPE, AND SELF-REGULATION AT FOUR TIME POINTS.....	221

FIGURE 18: CONDITIONAL GROWTH MODEL WITH A LATENT INTERCEPT, SLOPE, AND SELF-REGULATION AT FOUR TIME POINTS.....	223
FIGURE 19: AVERAGE CORSI BLOCKS GROWTH TRAJECTORY ACROSS TIME.....	227
FIGURE 20: AVERAGE HTKS GROWTH TRAJECTORY ACROSS TIME.....	228
FIGURE 21: AVERAGE DCCS GROWTH TRAJECTORY OVER TIME.....	230
FIGURE 22: AVERAGE COPY DESIGN GROWTH TRAJECTORY ACROSS TIME.....	231
FIGURE 23: AVERAGE PEG TAPPING GROWTH TRAJECTORY ACROSS TIME	233
FIGURE 24: AVERAGE SAR GROWTH TRAJECTORY ACROSS TIME.....	234
FIGURE 25: AVERAGE CFBRs GROWTH TRAJECTORY ACROSS TIME.....	236
FIGURE 26: TOOLS' IMPACT ON CHILDREN'S INITIAL EXECUTIVE FUNCTION STATUS AND GROWTH.....	240
FIGURE 27: TOOLS' EFFECT ON CHILDREN'S INITIAL CFBRs STATUS AND GROWTH.....	241
FIGURE 28: TOOLS' EFFECT ON CHILDREN'S INITIAL SAR STATUS AND GROWTH.....	242

SECTION I: INTRODUCTION

CHAPTER 1: Dissertation overview

Self-regulation predicts strong academic performance and reduced disciplinary problems for school-aged children (Lengua, 2003; Nota, Soresi, & Zimmerman, 2004). After the schooling years, robust self-regulation skills are linked with higher university completion (McClelland, Acock, Piccinin, Rhea, & Stallings, 2013), lower unemployment (Daly, Delaney, & Baumeister, 2015), improved body mass index (Schlam, Wilson, Shoda, & Mischel, 2013), and higher marital stability (Mischel, 2014). Thus, the benefits of self-regulation extend beyond the classroom and into many dimensions of life.

Given the extensive evidence regarding how self-regulation can positively affect people, this thesis seeks to explore the converse question – that is, how can people positively affect their own and others’ self-regulation? Specifically, this thesis will analyze an early childhood education curriculum called Tools of the Mind (Tools), which expressly aims to improve children’s self-regulation skills (Bodrova & Leong, 2007).

This DPhil is motivated by an imbalance in the literature: existing research demonstrates *why* cultivating self-regulation is important but not *how* to cultivate it. Teachers may want students to self-regulate but not know how to encourage students’ autonomous control. Does teacher-directed instruction compel children to self-regulate? Or do child-directed activities such as make-believe play enable students to manage their own impulses? Or can those strategies be effectively combined? This dissertation will investigate whether and how Tools promotes children’s self-regulation skills.

1.1 What is self-regulation?

Randi and Corno (2000) note that “there are almost as many definitions of self-regulated learning as there are lines of research on the topic” (p. 651). Indeed, self-regulation is a complex construct involving multiple psychological processes (Whiteside & Lynam, 2001), which has led to conceptual confusion in the literature (McClelland & Cameron, 2012). For example, a recent meta-analysis (Sitzmann & Ely, 2011) identified 16 underlying constructs that were all considered to be constituent elements of self-regulation (e.g., attention,

metacognition, persistence, help-seeking). The multi-faceted nature of the self-regulation construct is an example of the ‘jingle fallacy’ (Kelley, 1927), where one term signifies several interrelated but separable concepts.

Moreover, the literature exhibits conceptual overlap among terms such as self-regulation, executive function, effortful control, self-control, self-discipline, and various other terms. For example, Mischel, the researcher behind the famous marshmallow study (Mischel, Ebbesen, & Zeiss, 1972), refers to self-control, willpower, delay of gratification, and effortful control as interchangeable in his newest book (Mischel, 2014). This terminological interchangeability also surfaces in much of the published literature (Duckworth & Carlson, 2013; McClelland & Cameron, 2012), which exemplifies the ‘jangle fallacy’ (Kelley, 1927), whereby several terms signify the same underlying concept.

Although resolution of this jingle-jangle debate is outside the scope of this thesis, it is imperative to be clear about “what we mean” (Pring, 2012, p. 23) by self-regulation in order to properly analyze it. Thus, the section below outlines this thesis’ self-regulation definition, which should not be considered as a perfect definition but rather the definition that most aligns with this dissertation’s analytic aims.

1.1.1 Definition of self-regulation for this dissertation

Once again, this dissertation analyzes Tools of the Mind’s (Tools) curricular effects on children’s self-regulation. Given that analytic goal, only the self-regulatory components targeted by Tools will be investigated here. Specifically, the Tools’ website states that Tools is concerned with “the development of self-regulation/executive function in children” where “executive function and self-regulation are underlying mental processes that allow us to plan, intentionally focus, sustain or shift attention as needed, remember on purpose, and manage how much energy we put into a task” (Bodrova & Leong, 2015b).

Thus, based on the stated focus of the Tools program, self-regulation will be defined here as *one’s volitional control of attention, behavior, and executive function for the purpose of goal-directed action* (Blair & Ursache, 2011). The definition’s third element, executive function, comes with its own set of conceptual debates (Liew, 2012; Miyake et al., 2000). Nonetheless, executive

function is broadly agreed (David et al., 2003; A. Diamond, 2006; Fuhs, Farran, & Nesbitt, 2015; Röthlisberger & Neuenschwander, 2011) to contain three parts: 1) *inhibitory control* (suppressing impulses when those impulses are not aligned with desired goals), 2) *working memory* (remembering and manipulating information), and 3) *cognitive flexibility* (flexibly shifting attention across different tasks and sets of rules).

Whereas executive function refers to the cognitive capacities underlying thought and action, Schmitt et al. (2015) define self-regulation as the “integration of these three executive functions into overt behavior” (p. 21). That is, executive function skills refer to the cognitive processes that underlie action, whereas self-regulation also refers to the actual actions themselves. Overall, then, self-regulation refers to the autonomous control of the self by the self (Duckworth, 2011), which is underpinned by executive function processes.

Although some scholars also define self-regulation as including emotional regulation skills (Bronson, 2000; Liew, 2012; Portilla, Ballard, Adler, Boyce, & Obradović, 2014), this dissertation will focus most directly on the attentional and behavioral components. This decision derives from the stated focus of the Tools program, which explicitly targets the more cognitive and behavioral components of self-regulation.

Moreover, the *Handbook of child psychology* (Hyson, Copple, & Jones, 2006) considers emotion regulation to be “related to but distinct from the more cognitively focused self-regulation” (p. 21) and explains that there is “less emphasis in Tools of the Mind on emotional development” (p. 33). Thus, once again, the decision to focus this dissertation’s analyses on the cognitive aspects of self-regulation should not be construed as a claim that emotional regulation is irrelevant or unimportant; instead, the self-regulation definition in this paper is considered appropriate given the more cognitive focus of the Tools program (Bodrova & Leong, 2007).

In addition to emotion regulation, this dissertation will also not investigate related concepts such as creativity, self-efficacy, or motivation. Those elements, especially motivation, are conceptually distinguishable from self-regulation (Bronson, 2000). Instead, this dissertation will focus on students’ attentional and behavioral control in the Tools educational context.

1.1.2 Summary of self-regulation definition

Once again, this dissertation will proceed with the following self-regulation definition: *one's volitional control of attention, behavior, and executive functions for the purpose of goal-directed action* (Blair & Ursache, 2011). This definition focuses on the cognitive and behavioral aspects of self-regulation, which emerge as the foci of the Tools program.

The ideal of perfect self-regulation skills may be unattainable, especially for young children (Bandura, 1991). Given that Tools is aimed at preschool children, it may seem developmentally inappropriate to expect children to expertly self-regulate. Indeed, it is sensible to switch “from nouns to adverbs” (Pring, 2012, p. 29) to ask not whether a child self-regulates with complete *independence*, but, rather, whether he or she can be taught to self-regulate more *independently*.

Previous studies demonstrate that self-regulation is amenable to improvement (Barnett et al., 2008; Diamond, Barnett, Thomas, & Munro, 2007; Nunes et al., 2007) as well as deterioration (Karreman, Van Tuijl, & Marcel, 2006; Raver, Blair, & Willoughby, 2013). Consequently, it is crucial to identify education practices that foster self-regulation growth, which emerges as the central research rationale of this dissertation.

1.2 Research rationale

Developmental psychologists Posner and Rothbart assert (2000) that “understanding self-regulation is the single most crucial goal for advancing our understanding of development” (p. 427). The research community’s recent work has mirrored that assertion: between 1985 and 1990, 18 articles about executive function were published in peer-reviewed journals; between 2006 and 2010, the number of articles grew to 7,445 (Willoughby, Holochwost, Blanton, & Blair, 2014). According to a recent meta-analysis (Jacob & Parkinson, 2015), 11,000 articles pertaining to executive function were published between 2010 and 2015.

Despite the expansion of research in the field, the mechanisms underlying self-regulation development in educational contexts have yet to be adequately illuminated (Duncan et al., 2007; Fuhs, B. J. Zimmerman, 2008). For example, Fuhs et al. (2013) found that only 1.54% of variance in children’s executive function skills could be explained by the teacher to whom

the child had been assigned; that is, although all children improved their executive function skills over time, certain teachers did not produce substantial executive function gains for their students relative to other teachers, or else there would have been more systematic variance across classrooms. Fuhs et al. (2013) concluded that educators “do not yet know how to provide sufficient intentional instruction on CSR [cognitive self-regulation]” (p. 354).

This conclusion may be misguided, as it could have been the case that all teachers promoted children’s executive function skills equally well. Nevertheless, the results still suggest that very few, if any, teachers seem to have uncovered uniquely effective methods for cultivating self-regulation. Given self-regulation’s importance for school and life beyond, it is critical to identify educational practices that hone children’s self-regulatory skills. This dissertation seeks to address that very issue.

1.2.1 Why the focus on early childhood?

Because early self-regulatory competencies have been shown to predict later life outcomes (Moffitt, Arseneault, & Caspi, 2011; Sitzmann & Ely, 2011), early intervention in preschool contexts hold considerable promise for improving a child’s developmental trajectory. As Nobel laureate economist James Heckman noted, early “skill begets skill; learning begets learning” (Heckman & Masterov, 2007, p. 449). Consequently, relatively small self-regulatory differences in early childhood can be magnified to progressively larger differences over time (Alexander, Entwisle, & Kabbani, 2001; O’Shaughnessy, Lane, Gresham, & Beebe-Frankenberger, 2003).

Thus, early childhood emerges as an especially critical period in which to intervene. Previous longitudinal studies (Belfield, Nores, Barnett, & Schweinhart, 2006; Berrueta-Clement, 1984; Heckman & Masterov, 2007) have shown the transformative power of effective early educational interventions on subsequent life outcomes. Hence, this dissertation will focus on self-regulation development in students who experience Tools, which targets children between the ages of three and seven (Leong & Bodrova, 2011). Children of those ages often experience self-regulatory difficulties in many contemporary educational contexts, which will now be discussed below.

1.2.2 State of self-regulation in modern early childhood contexts

Research about the challenges of self-regulation promotion further underscores the need for improving early interventions. A U.S. national survey found that 46% of kindergarten teachers reported at least half of their students as routinely struggling with self-regulation (Rimm-Kaufman, Pianta, & Cox, 2000). In fact, American preschool students are three times more likely to be expelled for unmanageable behavior than primary and secondary students (Gilliam, 2005). Based on these statistics, it seems that many early childhood educational settings are neither meeting children's needs nor effectively promoting children's self-regulation skills.

Certain subpopulations of children face unique self-regulation challenges from a young age. Children growing up in poverty are more likely to experience self-regulatory problems (Raver et al., 2013; Raver, 2012), which make low-income children susceptible to disciplinary action (Alloway, Lawrence, & Rodger, 2013; Miller, Nevado-Montenegro, & Hinshaw, 2012). For example, a Washington DC city report (Office of the State Superintendent of Education, 2013) revealed that three- and four-year-old students received 181 suspensions during the 2012-2013 school year, most of which were given to students in low-income schools.

Moreover, many children have been diagnosed with attention deficit hyperactivity disorder (ADHD) and conduct disorder, where children exhibit chronic difficulties in regulating their attention and behavior. In 2013, 11% of American children between the ages of 4 and 17 had been diagnosed with ADHD, which reflects a 41% increase in diagnoses over a single decade (Center for Disease Control, 2013). In the UK, 7% of British boys and 3% of British girls aged 5 to 10 meet the diagnostic criteria for conduct disorder (National Institute for Health and Care Excellence, 2013), which presents challenges to the educators responsible for student learning (Webster-Stratton, Reid, & Stoolmiller, 2009).

Of course, the observed increase in self-regulatory issues may reflect more about these national systems' problems than the children's problems. That is, the high expulsion rates for preschool students in the United States (Gilliam, 2005), the rising conduct disorder incidence in the United Kingdom (National Institute for Health and Care Excellence, 2013), and the growing reports of self-regulation issues across Europe and Asia (Ben-Ari, 1995;

Kwon, 2003) may simply reflect societal shifts rather than genuine increases in children's self-regulatory issues.

Nevertheless, given the benefits of robust self-regulation skills for children and for the adults they will become, it is important to identify educational methods that cultivate all children's self-regulation. Fortunately, the number of self-regulation interventions has increased alongside the rising concerns regarding children's self-regulation issues (Harris, Friedlander, & Graham, 2005; Soares, Vannest, & Harrison, 2009; Thompson, Ruhr, Maynard, Pelts, & Bowen, 2013), especially for children with special educational needs (SEN; Gulchak, 2008; K. Jones, Daley, Hutchings, Bywater, & Eames, 2007).

In contrast to the rich literature regarding interventions for SEN children, the body of evidence on self-regulation-oriented programs for mainstream student populations is sparse. Given the relative dearth of self-regulatory programs for mainstream children, this dissertation will not analyze interventions designed specifically for SEN students. Instead, this thesis will analyze Tools, which, to the best of my knowledge, is the only mainstream and comprehensive preschool curriculum to identify self-regulation cultivation as its paramount aim (Bodrova & Leong, 2007).

1.3 Research focus: Tools of the Mind

Tools is comprised of “activities with a dual purpose – to develop foundational executive function skills and self-regulation skills, at the same time they develop core academic skills” (Bodrova & Leong, 2015b). Many existing self-regulation interventions involve either individualized plans for specific children (Gulchak, 2008; Soares et al., 2009) or a set of exercises to supplement an existing curriculum (K. L. Bierman, Domitrovich, Blair, Nelson, & Gill, 2008; Domitrovich, Cortes, & Greenberg, 2007); by contrast, Tools intends to be a comprehensive curriculum that covers all subjects and is delivered to all students in a mainstream classroom (Bodrova & Leong, 2007).

Tools derives its inspiration from Vygotsky, who, in *Thought and language* (1962), develops the concept of “mental tools” (p. 164). Mental tools aim to extend mental faculties in the way that physical tools extend physical faculties. For example, although young children typically

struggle with task focus, they can be taught to use private speech (i.e., speech intended to guide one's own behavior as opposed to communicating with others) in order to maintain concentration. In this case, private speech serves as a mental tool that enables children to direct their attention amid distraction (Vygotsky, 1962).

The Tools curriculum aims to bring Vygotsky's ideas into practice by honing children's self-regulation abilities with a set of mental tools. According to the Tools curriculum developers, all parts of the school day contribute to students' self-regulation growth: "The Tools program has designed the daily schedule, the way in which teachers manage transitions [. . .], and all classroom activities to support the development of self-regulation and executive function skills" (Bodrova & Leong, 2015b). Specifically, Tools includes over 60 activities that simultaneously target students' self-regulation and academic skills.

Although all parts of the school day contribute to self-regulation development, the Tools curricular developers assert that "the best way for children to practice self-regulatory behaviors is to engage in mature make-believe play" (Bodrova & Leong, 2015a). Given that assertion, the Tools manual (Leong & Bodrova, 2011) directs teachers to implement the Tools make-believe play block every day for up to one hour (p. 217). The connection between the make-believe play block and self-regulation is highly specific and will be described in section 2.4.1.

One Tools activity emblematic of Tools' dual focus on self-regulation and academic skills is called buddy reading (Leong & Bodrova, 2011, p. 342). Buddy reading involves two students who jointly read a book. One child receives a picture of a mouth, which designates him or her as the reader; the other child receives a picture of an ear, which designates him or her as the listener. The reader then reads the book while the other child actively listens and checks for decoding errors. The children then exchange pictures and roles after the first reader completes the book (Leong & Bodrova, 2011).

Given proper execution, buddy reading simultaneously targets literacy and self-regulation. Since self-regulation requires executive function in the service of desired goals, buddy reading should theoretically hone all three parts of executive function: 1) *working memory*, as children remember and act out either the listening or reading role), 2) *cognitive flexibility*, as

children flexibly switch across the two roles when the reader finishes the book, and 3) *inhibitory control*, as children suppress impulses to switch roles at inappropriate times (e.g., the listener trying to become the reader before his or her turn).

In sum, whether children are engaged in literacy, math, or any part of the daily schedule, each Tools activity aims to target self-regulation. Tools is designed to be implemented by classroom teachers throughout a full academic year (Leong & Bodrova, 2011). Moreover, in contrast to programs that target only children with self-regulatory deficits, the Tools curriculum developers argue that self-regulation instruction “should not be reserved only for ‘problem’ children” and that “all children benefit from practicing deliberate and purposive behaviors” (Bodrova & Leong, 2005, p. 35). Thus, Tools’ comprehensive nature emerges as a key mechanism of its purported efficacy in improving children’s self-regulation (Blair & Raver, 2015).

1.4 Policy context: Why is it important to evaluate Tools?

Since its development in 1993, Tools has been increasingly implemented in parts of the United States, Canada, and South America (Blair & Raver, 2014). One illustrative case study of Tools’ proliferation is Washington DC, where Tools expanded from a two-school pilot in 2010 to a 28-school program in 2011 at a \$1.5 million cost beyond the city’s traditional curriculum budget (Turque, 2011). As of the 2016-2017 school year, 38 out of the city’s 60 preschools had implemented Tools (District of Columbia Public Schools, 2016).

Beyond Washington DC, various schools in 20 U.S. states, three Canadian cities, and certain regions of Chile are currently using the curriculum (Bodrova & Leong, 2015a). In the face of Tools’ proliferation, the question emerges: does the program significantly improve children’s self-regulation?

Despite the curriculum’s growth in popularity, the evidence base behind it remains unclear (Jacob & Parkinson, 2015). Given that the curriculum developers claim that the program improves children’s self-regulation (Bodrova & Leong, 2007), and given that the Tools costs \$3000 per classroom in the first year alone for professional development and materials

(United States Department of Education, 2008), it is critical to evaluate whether the program achieves its aims.

If Tools does not deliver on its stated outcomes, then this expansion in North and South America, and perhaps soon to Europe (Highfield, 2007), may be misguided. By contrast, if Tools does deliver on its stated outcomes, then it should be recognized as a program that improves children's self-regulation, and, thus, improves children's lives.

1.5 Theoretical frameworks for this dissertation

This dissertation contains two distinct theoretical frameworks: the first undergirds the Tools curriculum and the second undergirds this dissertation's analytic approach. The first, Vygotsky's (1962) cultural-historical theory, was used by the Tools developers (Bodrova & Leong, 2007) to inform Tools' instructional approach through a focus on self-regulation. Chapter Two of this dissertation more thoroughly explicates the cultural-historical theory and its application to Tools.

The second theoretical framework is Bronfenbrenner's bioecological model, which underpins the analytical and statistical portions of the dissertation. Specifically, the bioecological model posits that children are affected by and nested within progressively larger systems. Given that all analyses in this dissertation involve multilevel modeling, where smaller units are nested within larger units, the bioecological model emerges as a suitable theoretical basis for this dissertation's statistical approach. Chapter Three of this dissertation more thoroughly explicates the bioecological model and its application to these analyses.

1.6 Structure of the analyses in this dissertation

This dissertation includes two empirical studies: 1) a research synthesis (i.e., a systematic review and meta-analysis) of existing Tools evaluation research, and 2) an analysis of the specific activities that collectively comprise Tools. Based on gaps in the existing literature (see section 7.5.1), the research aims and questions of each study are as follows:

Study one: Research synthesis of existing Tools studies

Study one aims to identify, appraise, and synthesize existing evidence regarding Tools in order to evaluate Tools' effectiveness in promoting children's self-regulation skills. In order to do so, I will conduct the first Tools research synthesis study, which includes both a systematic review and multilevel meta-analysis (see section 8.4). The research questions for study one include the following:

- What is Tools' aggregate effect size observed across the existing studies?
- Is that aggregate effect size significantly heterogeneous? That is, does Tools' effectiveness vary significantly across the existing studies?
- Can heterogeneity in the aggregate effect size be explained by child-level and study-level characteristics? For example, do child characteristics (e.g., free school meal status, gender) explain why Tools was effective in some contexts and not others?

Study two: The secondary data analysis of Tools and its activities

Study two analyzes the associations between Tools activities and children's self-regulation development using secondary data from 80 American preschool classrooms. Study two has two main aims. Firstly, study two intends to replicate findings from a previous study (Farran & Wilson, 2014), which used the same dataset as was analyzed for study two of this thesis.

Although study two uses the same dataset as Farran and Wilson (2014), study two goes beyond Farran and Wilson's (2014) research in two key ways. Firstly, study two employs latent variable structural equation modeling instead of traditional multilevel modeling. Whereas Farran and Wilson (2014) did not find positive impacts for Tools on children's self-regulation, the self-regulation measures they used contained measurement error, which could have distorted the results. The analysis in this thesis accounts for measurement error using latent variable structural equation models to more accurately estimate Tools' effect.

Secondly, study two goes beyond Farran and Wilson (2014), as well as all existing Tools research, by being the first research to analyze specific Tools activities. Whereas previous studies have investigated Tools' effectiveness as an entire curriculum, no studies have

analyzed the specific instructional activities that collectively comprise the program. This activity analysis is the central aim of study two. Based on literature gaps that will be described in Chapter Seven, study two addresses the research questions presented below:

1. How does children's self-regulation develop, on average, between the start of pre-kindergarten and the end of first grade?
2. Does Tools differentially affect children's self-regulation developmental trajectories vis-à-vis comparison curricula?
3. Which, if any, subject-specific groupings of Tools activities (e.g., literacy, make-believe play, math) predict children's self-regulation trajectories?
4. Which, if any, of the 61 Tools instructional activities predict children's self-regulation trajectories?

Those questions, as well as the research hypotheses for each, will be discussed more thoroughly in Chapter Eleven. The research hypotheses were informed based on existing self-regulation literature, which will be extensively described in the literature review of this dissertation (Section II). Before moving onto the literature review, section 1.7 below outlines this dissertation's structure to provide clear expectations for what is to come.

1.7 Overall dissertation structure

This dissertation contains 14 chapters distributed among five sections: 1) Introduction, 2) Literature review, 3) Methods, results, and discussion for study one (i.e., the systematic review and meta-analysis), 4) Methods, results, and discussion for study two (i.e., the analysis of specific Tools activities), and 5) Conclusion. Each of the five sections is described in more detail below.

1.7.1 Section I: Introduction

The introduction section contains the present dissertation overview chapter as well as the two theory chapters (Chapters Two and Three). The first theory chapter, Chapter Two, outlines Vygotsky's socio-cultural theory and explains how that theory undergirds Tools. The second theory chapter, Chapter Three, covers Bronfenbrenner's bioecological model and explains its relevance for this dissertation's statistical analyses.

1.7.2 Section II: Literature review

The literature review section includes five chapters pertaining to the following topics:

A review of self-regulation's correlates at multiple stages of the life course

Chapter Four presents previous research about self-regulation's associations with academic, social, behavioral, and other life outcomes. That is, why should we care about children's self-regulation development? And, by extension, why should we care about evaluating Tools' curricular effects? This chapter aims to provide motivation for researching Tools as a mechanism to improve children's self-regulation.

A review of the biological and contextual predictors of self-regulation development

While this dissertation investigates Tools' impact on self-regulation, other influences on children's self-regulation are also important to understand. Chapter Five assesses the evidence base for several person-level and contextual correlates of self-regulation, including genetics, neurobiology, poverty, parenting, and education.

A review of early childhood interventions that target children's self-regulation

Chapter Six presents research regarding a set of early childhood self-regulation interventions. All of the programs reviewed in Chapter Six are supplements that are designed to augment existing early childhood curricula.

A review of educational curricula that affect children's self-regulation

Chapter Seven focuses on a set of comprehensive early childhood curricula (as opposed to the supplemental interventions described in Chapter Six) with an emphasis on their impacts on children's self-regulation development. Given that this dissertation analyzes an early childhood curriculum, it is important to provide background research on other curricular options in the field. The chapter concludes with a brief discussion of Tools, especially regarding the literature gaps in the Tools evidence base that this dissertation seeks to fill.

1.7.3 Section III: Methods, results, and discussion for study one

Subsequent to the literature review, Section III presents the research synthesis for study one, which includes both a systematic review and a multilevel meta-analysis. That is, according to existing evaluation studies, does Tools effectively improve children's self-regulation vis-à-vis comparison curricula? Section III includes three chapters:

A methodology chapter

Chapter Eight outlines the systematic review and multilevel meta-analysis methods used to address the research questions for study one.

A results chapter

The results chapter (Chapter Nine) assesses Tools' impact on children's self-regulation as well as academic achievement in literacy and math. This chapter takes the form of a systematic review and meta-analysis.

A discussion chapter

The discussion chapter contextualizes the findings from study one within the existing literature base. This chapter also identifies study one's strengths and limitations before providing recommendations for further research.

1.7.4 Section IV: Methods, results, and discussion for study two

After presenting study one in Section III, study two will be presented in Section IV. Once again, study two uses secondary data on 1145 children in 80 classrooms to analyze which, if any, Tools activities predict children's self-regulation growth trajectories. As with Section III, Section IV includes three chapters:

A methodology chapter

Chapter Eleven outlines the secondary data sample, ethics, and multilevel structural equation modeling techniques to address the research questions of study two.

A results chapter

Chapter Twelve presents the findings regarding children's self-regulation growth trajectories, and whether, overall, Tools affects self-regulation differently than the comparator curricula. The existing study to use this same dataset (Farran & Wilson, 2014) indicated mostly null and negative effects for Tools on self-regulation outcomes over time. However, their study did not account for measurement error in the self-regulation constructs nor did they analyze the full sample of children (see section 12.2 for more information). Thus, the first part this results chapter seeks to replicate those findings using more rigorous statistical methods.

The second part of Chapter Twelve analyzes each individual Tools activity as well as subject-specific blocks of Tools activities (e.g., literacy, math, make-believe play). The goal of these analyses is to identify specific educational practices that promote children's self-regulation, some of which could potentially be replicated beyond the Tools curricular context.

A discussion chapter

Chapter Thirteen contextualizes the findings from study two within the existing literature. This chapter also identifies study two's strengths and limitations before providing recommendations for further research.

1.7.5 Section V: Conclusion

This dissertation's fifth and final section will be a brief conclusion chapter that synthesizes findings from study one and study two before explaining the findings' relevance for policy and practice.

CHAPTER 2: Vygotsky, the cultural-historical theory, and Tools of the Mind

This chapter presents Tools' Vygotskian theoretical underpinnings, its theory of change, and its pedagogical approach. The Tools curriculum developers, Russian psychologist Elena Bodrova and American psychologist Deborah Leong, emphasize (2007) the importance of understanding Vygotskian theory in order to fully understand the curriculum. Thus, the following sections provide information about Vygotsky's research in order to illustrate how Tools developed from his work.

2.1 Vygotsky's research contribution

The contrast between the length of Vygotsky's life and the depth of his research contribution is remarkable: in 37 years of living, Vygotsky published over 200 articles and books¹ about human development, psychology, education, and other related topics (van der Veer, 2007). Despite the eclecticism of his research interests, he is arguably most famous for his cultural-historical (also known as 'socio-cultural') theory (van der Veer, 2007).

Although the term 'cultural-historical' never explicitly appears in Vygotsky's writings, it is used to identify Vygotsky's overarching philosophy, which was shared by some of his contemporaries such as psychologists Luria and Mikhailov (Bronson, 2000; Shayer, 2003). The sections below describe the cultural-historical theory, connect it to Vygotsky's focus on self-regulation, and then explain Vygotsky's influence on the Tools program.

2.2 Cultural-historical theory

The cultural-historical theory posits that children acquire knowledge through interaction with other members of their culture. According to Vygotsky (1962), cultures develop

¹ Most of Vygotsky's work was not translated into English until decades after his death in 1934. In fact, between 1969 and 1986, 82% of the English-language citations to Vygotsky were to either the 1962 English translation of *Thought and Language* or the 1978 translation of *Mind and Society* (Gredler, 2009). It was not until the following year, 1987, that a new translation of six Vygotskian lectures was published, which was then followed by several further releases of previously untranslated work in the 1990s. Thus, many of the Vygotsky citations below have a publication date in the 1990s; this was the first time that much of his work had been translated for wider consumption.

cultural tools (e.g., language, number systems) that enable members of that culture to address shared problems (e.g., how to communicate with other members of society).

Thus, the theory's 'culture' element signifies that cultures develop shared systems to address problems, whereas the theory's 'historical' element refers to the evolution and transmission of those cultural tools across generations. Specifically, adults within a culture transmit the tools to children, who then continuously transmit that cultural knowledge among one another through social interaction (van der Veer, 2007).

To Vygotsky, schools play a critical role in enabling members of the culture to meaningfully interact with one another, which then enables "man as a social type" to develop out of "man as a biological type" (Vygotsky, 1930, p. 58). Although Vygotsky argued that schools should facilitate cultural exchange among students and teachers, Vygotsky's writings did not provide specific prescriptions as to how this would manifest in classroom settings. Rather, as the biographer van der Veer writes (2007), Vygotsky was interested in more fundamental questions of human development:

Vygotsky did not intend to develop concrete ideas about the organization of schools, the development of curricula, and so on. Rather, in a sense, what he intended to do was to develop a theory of what it means to become a human person. What makes us human beings different from other animals? What is human development about? [. . .] These were the fundamental questions that bothered Vygotsky and which he tried to answer with his cultural-historical theory of the higher mental functions. (p. 139)

The "higher mental functions" to which van der Veer refers are what Vygotsky believed to distinguish human beings from other animals (Vygotsky, 1981). Nearly all animal species have elementary mental functions, which include basic capacities such as sensation and sensorimotor control (Fernyhough, 2009; Fox & Riconscente, 2008; Vygotsky, 1981). Human beings, by contrast, are the exclusive possessors of higher mental functions, which include voluntary attention, working memory, and conceptual thinking (Vygotsky, 1998a).

Thus, Vygotsky's notion of 'higher mental functions' has substantial conceptual overlap with self-regulation, which also refers to voluntary attention, memory, and cognitive control. To Vygotsky, children's development of higher mental functions ranked as "the crowning achievement of cognitive development" (Vygotsky, 1998a, p. 87). Thus, the section below connects Vygotsky's higher mental functions concept with self-regulation and explains how self-regulatory functions develop within the cultural-historical context.

Vygotsky on higher mental functions and self-regulation

Vygotsky never used the term self-regulation, which did not consistently appear in the literature until over thirty years after his death, but, again, the concept of self-regulation is consistent with Vygotsky's notion of higher mental functions (Fernyhough, 2009). For example, Vygotsky argued that a two-year-old toddler initially "does only what surrounding objects nudge him to do" (Vygotsky, 1998, p. 243), whereas higher mental functions enable children to become "masters of their own behavior" (Vygotsky, 1962, p. 147). That is, infants reactively respond to their environment in a similar way that other animals do – their elementary mental functions do not involve conscious deliberation or volition.

As children grow older and interact with members of their culture, children become less reactive and instead develop higher mental functions such as voluntary attention and autonomous behavioral control. According to Vygotsky (1997a), "the essence of cultural development [. . .] consists of man mastering processes of his own behavior" (p. 242-243). Thus, the core of the cultural-historical theory rests on the idea that children learn to regulate their behavior so that they can undergo cultural development.

As Vygotsky's (1998) quote about the two-year-old toddler doing "only what surrounding objects nudge him to do" (p. 243) indicates, Vygotsky asserted that young children are not born with higher mental functions. In fact, Vygotsky wrote that full self-regulation could not manifest until adolescence (Vygotsky, 1998b, p. 190). Although children cannot fully develop self-regulation, van der Veer (2007) translated from Vygotsky's original Russian to explain Vygotsky's focus on education as a mechanism to cultivate children's cognitive control in primary school:

The major products of instruction in elementary school are *osoznanie* and *ovladenie*. *Osoznanie* can be translated as "becoming conscious of something" or "conscious realization." The child becomes capable of reflecting about his or her own actions and utterances. *Ovladenie* can be translated as "mastery" or "control." Because the child consciously realizes what he or she is doing, he or she can deliberately carry out actions or operations. (van der Veer, 2007, p. 90)

Thus, Vygotsky argued that education's central aim was to make children capable of controlling (*ovladenie*) their mental processes (*osoznanie*). The critical question remains how to accomplish such a goal within a school. Specifically, what educational practices can hone children's cognitive control skills? Vygotsky's writings include at least three ideas that bear

on that question: mediation, the zone of proximal development (ZPD), and the importance of make-believe play. Each of these concepts has a central role in the Tools program, so each is discussed in turn below.

2.2.1 Mediation

Mediation refers to the use of concrete symbols to abstractly represent a person, object, or concept (Vygotsky, 1997). The mediator itself refers to the symbol used to represent the object, person, or concept (e.g., the written word ‘cat’ to represent an actual cat). Although these mediators are first external and unintelligible to the child, the child then learns to use and internalize the mediators “as a means of directing and mastering mental processes” (Vygotsky, 1997a, p. 126).

Vygotsky also referred to mediators as “tools” (Vygotsky, 1997a, p. 85), which explains the nominal derivation of the Tools curriculum. Like physical tools, these mental tools enable children to extend beyond their current capacity and progress toward the higher mental functions they have not yet mastered. Specifically, Vygotsky (1997a) wrote that “the psychological tool modifies the entire course and structure of mental functions” (p. 85) and that a child learns “mastery of one’s own behavior with the assistance of symbolic stimuli” (Vygotsky & Luria, 1994, p. 135).

Thus, mediators assist children in controlling their higher mental functions. In a sense, mediators serve as scaffolds that support learners until the child has internalized a skill or concept; at that point, the mediator can be discarded, and the child can independently direct his or her own thinking without the external mediator.

An example of a mediator in the Tools context is the buddy reading cards described in section 1.3; the physical cards remind children of their roles and keep them focused, which would be more difficult for the children if they did not have cards. The gap between that which children can do independently and that which they can do with external supports (e.g., the buddy reading cards, teachers) represents another central Vygotskian concept: the zone of proximal development (ZPD).

2.2.2 Zone of proximal development

Whereas mediators are objects or concepts that support children's learning, Vygotsky argued that children's capacities also evolve through partnerships with more knowledgeable others (e.g., teachers and peers). This notion forms the foundation for Vygotsky's zone of proximal development (ZPD) concept. Given the centrality of ZPD in Vygotsky's educational philosophy, I have included Vygotsky's (1933a) original explication of the idea:

The zone of proximal development of the child is the distance between the level of his actual development, established with the help of problems independently solved, and the level of the child's possible development, established with the help of problems solved by the child under the guidance of an adult or in cooperation with his more intelligent partners. The zone of proximal development refers to functions that have not yet matured, but are in the process of maturing, functions that mature tomorrow, that are now still in their embryonic form; functions that cannot be called the fruits of development, but the buds of development, the flowers of development. (p. 7)

Thus, to Vygotsky, an accurate assessment of a child's skills must consider both the child's ability when operating alone as well as when operating with a more competent other. Those two points (i.e., independent versus collaborative work) delineate the boundaries of a child's zone of proximal development.

Vygotsky further explained that higher mental functions “develop from a shared function to an individual function” (Vygotsky, 1987, p. 21); this suggests that teachers and peers must first help a child control his or her behavior before the child can do so independently. Moreover, in all learning domains, Vygotsky's theory directs teachers work with children on learning goals that are slightly beyond children's current individual abilities so that children may grow more quickly (Bodrova & Leong, 2007).

The notion that instruction can be used to accelerate development is a theme to which Vygotsky repeatedly returned. Vygotsky asserted that not only *can* learning lead development but also that certain forms of development *can only* occur given appropriate instruction: “Only that instruction is good that runs ahead of development [. . .] The correctly organized instruction of the child calls into life a whole series of developmental processes that without instruction would have been altogether impossible” (Vygotsky, 1933, p. 15-16). One such instructional technique that Vygotsky emphasized for propelling children's development forward was make-believe play, which will now be described in the section below.

2.2.3 Make-believe play

Vygotsky was among the first theorists to propose that make-believe play fosters children's learning and control over themselves² (Berk & Meyers, 2013, p. 98). In order to be effective, Vygotsky argued (1933b) that play scenarios had to exhibit a certain structure. Specifically, effective play scenarios require three elements: 1) children determine an imaginary scenario, 2) they negotiate roles for themselves and one another, and 3) they act out those roles with fidelity (i.e., not switch or cease a role simply because one has lost interest in it).

According to Vygotsky (1933b), any play scenario that includes those three elements hones children's self-regulation (or, in Vygotsky's terms, higher mental functions) more than any other activity: "At every step the child is faced with a conflict between the rule of the game and what he would do if he could suddenly act spontaneously. In the game he acts counter to what he wants . . . [achieving] the maximum display of willpower" (1933b, p. 14).

Specifically, make-believe play requires children to focus on a role (e.g., a grocer), enact that role (e.g., put food into bags), and inhibit the impulse to suddenly switch roles (e.g., become the store manager during a play scenario), even when the child wishes to act spontaneously.

Vygotsky (1933b) also wrote that children exhibit their "greatest self-control in play" (p. 13), which highlights his focus on make-believe play as the optimal mechanism for cultivating self-regulation. In addition to self-regulation development, Vygotsky also argued that make-believe play scenarios compel a child to behave "beyond his average age, above his daily behavior; in play it is as though he were a head taller than himself" (1933, p. 12). The idea of achieving beyond one's limitations overlaps substantively with Vygotsky's mediation and ZPD concepts, which collectively form Tools' theoretical foundation as described below.

² Vygotsky's focus on pretend play may have been influenced by his personal fascination with theater. Vygotsky routinely attended Shakespearean theater in Russia, and he wrote his master's thesis about *Hamlet* (van der Veer, 2007). Although there is no published research about this as of yet, the overlap between Vygotsky's love of theater and the paramount importance of make-believe play in his developmental theory seems worth noting.

2.3 Vygotsky's influence on Tools of the Mind

When considering Vygotsky's role in the creation of Tools, the Tools curriculum developers write that Tools "is inspired by the work of the Russian psychologist Lev Vygotsky and his students, and at the same time, is rooted in cutting edge neuropsychological research on the development of self-regulation/executive function in children" (Bodrova & Leong, 2015b).

Similarly, according to the *Handbook of child psychology* (Hyson et al., 2006), "Tools of the Mind is one of the most theory-driven curricula currently being implemented in early childhood education" (p. 37). Tools' focus on Vygotskian theory is further underscored by the curricular developers' book title, which is *Tools of the Mind: The Vygotskian approach to early childhood education* (Bodrova & Leong, 2007).

Given that Vygotsky's research contribution was both expansive and eclectic (section 2.1), it is important to reiterate that the Tools developers were specifically interested (Bodrova & Leong, 2007) in Vygotsky's notion of the higher mental functions. In line with other Vygotskian theorists (Ferryhough, 2009; Fox & Riconscente, 2008; Shayer, 2003), the Tools developers explicitly connect the higher mental functions concept with self-regulation: "For Vygotskians, higher mental functions are self-regulated, mediated, and learned mental functions, which makes them central to the discussion of the development of self-regulation" (Bodrova & Leong, 2011, p. 12).

As mentioned in section 1.3, Tools embeds self-regulation within all of its activities. As with Vygotsky, the Tools curricular developers maintain a central focus on make-believe play in the Tools program. In fact, they assert that "the best way for children to practice self-regulatory behaviors is to engage in mature make-believe play" (Bodrova & Leong, 2015b), where 'mature' refers to the three characteristics of play enumerated in section 2.2.3.

Although the Tools developers designed the curricular activities based on Vygotskian theory, they acknowledge that Vygotsky's theory does not "provide recipes for any classroom situation" (Bodrova & Leong, 2007, p. 17). Nonetheless, they developed Tools in order to capture Vygotsky's philosophy using the curricular theory of change described below.

2.4 Tools' theory of change

Tools' theory of change contains three stages: 1) the teacher regulates the students, 2) the students regulate one another, and 3) the students self-regulate (Bodrova & Leong, 2007). When the students first arrive in a classroom, Vygotsky wrote (1962) that they are “slaves to their environment,” and education’s aim must be to transform them into “masters of their own behavior” (p. 147). The Tools curricular developers attempted to harness Vygotsky’s philosophy through the teacher-regulated, other-regulated, and self-regulated theory of change model depicted in figure 1.

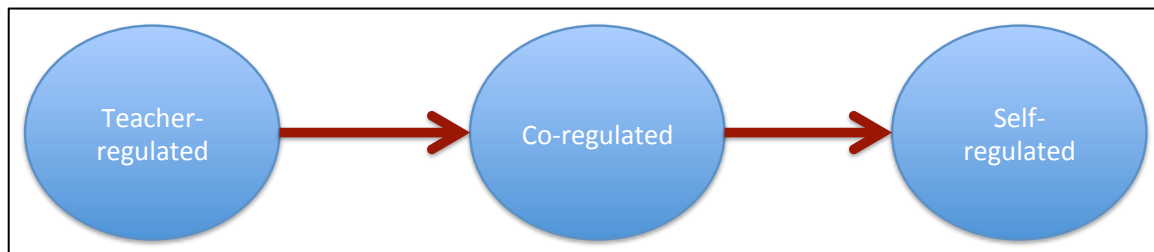


Figure 1: Tools theory of change (adapted from Bodrova & Leong, 2007)

Vygotsky wrote that “inner regulation of purposeful activity originates in external regulation” (Vygotsky & Luria, 1994, p. 164). In other words, a child’s ability to regulate his internal thoughts and actions must begin with someone outside of the child (i.e., an adult or more competent peer) who first regulates the child’s behavior. In one of his earlier writings, Vygotsky refers to the external adult as “the ideal form” of cognitive control who interacts with “the present form” (Vygotsky, 1994, p. 366), or the child. Over time, the child’s ‘present form’ evolves until it hopefully reaches the ‘ideal form’ of cognitive control.

Along a similar line of reasoning, Bruner coined the term ‘scaffolding’ (Wood, Bruner, & Ross, 1976), which refers to a teacher’s changing levels of support for children in pursuit of their learning goals. As a child becomes more competent, the teacher gradually removes support until the child can work independently.

This same process applies for the development of self-regulation. In the words of Vygotsky's colleague, Luria, "the source of the volitional act is the child's communication with adults" (Luria, 1981, p. 89). That is, Luria argues that a child's ability to autonomously self-regulate is first rooted in a child's interaction with the adults around the child. The Tools developers refer to Vygotsky's, Luria's, and Bruner's theories when expounding upon Tools' theory of change (Bodrova et al., 2011).

Following teacher-regulation, the Tools developers (2011) identify children's co-regulation as the second part of the theory of change: "To develop self-regulation, children need to have an opportunity to engage in other-regulation. Other-regulation implies that children act both as subjects of another person's regulatory behaviors and as actors regulating another person's behaviors" (p. 73). This concept derives directly from Vygotsky's book *Mind and society* (1978), where he argued that children's higher mental functions are strengthened through social interactions with other children before eventually becoming internalized and autonomously exercised later on.

In sum, the Tools theory of change specifies children's self-regulation as the ultimate goal and then identifies teacher-regulation and co-regulation as the steps toward that end. The Tools logic model aligns well with Vygotsky's claim that "inner regulation of purposeful activity originates in external regulation" (Vygotsky & Luria, 1994, p. 164). Similarly, the Tools developers' claim that "learning to control impulsive behavior is the most critical requirement for young children" (Bodrova & Leong, 2005, p. 30) reflects Vygotsky's notion that children's main developmental goal is to become "masters of their own behavior" (Vygotsky, 1962, p. 147). That prioritization of self-regulation is made manifest through the Tools curricular activities, four of which will be described below.

2.4.1 Four Tools activities emblematic of Vygotskian theory

The sections below describe four Tools activities that represent the Tools developers' attempt to apply Vygotskian theory to early childhood educational practice. The first, make-believe play, is identified by the Tools developers as the critical driver of self-regulation growth (Bodrova & Leong, 2007) and is thus a cornerstone of the curriculum.

Beyond make-believe play, though, I should note that neither the curricular developers nor the Tools curriculum manual identified the other three activities below as especially emblematic of Tools; rather, I chose these activities because they illustrate how Tools combines self-regulation and academic skills into each activity. In addition to the four activities below, the remaining Tools activities are briefly described in Chapter Twelve.

Tools activity #1: Make-believe play

Among the original Tools activities is the make-believe play block, which is meant to occur every day in Tools classrooms for up to one hour (Bodrova & Leong, 2013). As explained in section 2.3, Vygotsky argued (1933b) that effective play scenarios involve three elements: 1) children determine an imaginary scenario, 2) they negotiate roles for themselves and one another, and 3) they act out those roles with fidelity (i.e., not switch or cease a role simply because one has lost interest in it). In order to establish such structured play scenarios, teachers work with children to create play plans as depicted in figure 2.



Figure 2: Tools make-believe play plan (Bodrova & Leong, 2007)

For the make-believe play block, the Tools manual (Leong & Bodrova, 2011) first directs teachers to convene a small group of students who collectively determine a play scenario. Secondly, children negotiate roles to which each child will adhere throughout the play block (i.e., children cannot switch roles in the middle). Each child then creates a play plan

including his or her name, a picture that illustrates an individual child's play plan, and then a textual description of that plan. Brenda's plan from figure 2 indicates that she will be Sleeping Beauty and marry a prince.

Thus, make-believe play involves writing practice, drawing practice, and goal-oriented thinking to guide children's subsequent behavior. If children forget their roles, then the teacher and/or other children can reference the play plans (Leong & Bodrova, 2011). Thus, the play plan constitutes a mediator that directs the child's play behavior. The play-planning process gives way to the actual play scenario, which is where Vygotsky argues (1967, p. 102) that children's self-regulation is maximally taxed.

Tools activity #2: Scaffolded writing (literacy)

Another central Tools activity is scaffolded writing (Bodrova & Leong, 2001), which aligns with Vygotsky's mediation and ZPD concepts. During this activity, children articulate a sentence that they would like to write but cannot, count the number of words, and then observe as the teacher draws a line for each word in the sentence (see figure 3).

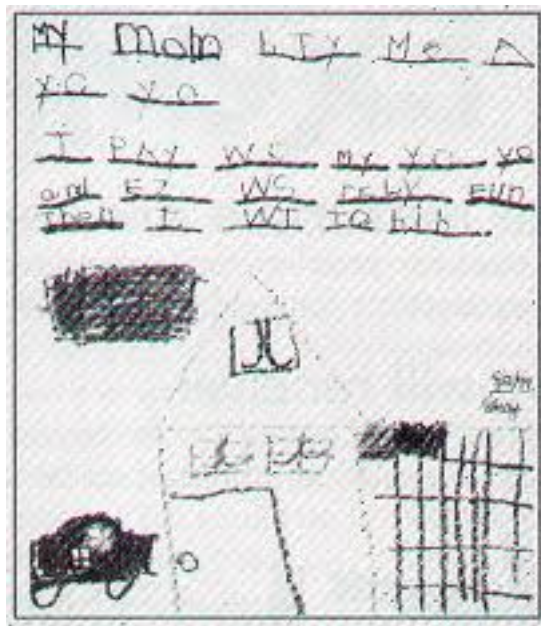


Figure 3: Scaffolded writing example (Bodrova & Leong, 2007)

As figure 3 shows, the lines serve as mediators to represent words that the children are not yet able to independently write. Scaffolded writing exemplifies the ZPD concept because the activity contains elements that the children both *can* and *cannot* independently complete. That is, children are able to independently say the sentence aloud and count the number of words with the teacher; however, the children are not yet able to write the full sentence independently, which is where the teachers offer individualized support.

The space between children's writing abilities on their own versus their abilities with the teacher's support delineates the ZPD's boundaries, which are different for each child. For example, some students need help simply identifying each word's initial sound, whereas others only need help with difficult sounds (e.g., phoneme-grapheme incongruities such as 'sh' or 'th'). In either circumstance, the teacher supports the child slightly beyond his or her independent abilities using mediators that the child, once ready, will eventually abandon.

Tools activity #3: Numerals game (mathematics)

Numerals game requires a pair of children to use cards with visual symbols (e.g., cars, bugs) and perform the target operation depicted on the card. For example, children learning addition could go through a set of cards where they have to add quantities of animals. Much like the buddy reading activity (see section 1.3), children receive either a 'counter' card or a 'checker' card to designate their personal roles (Leong & Bodrova, 2011, p. 470).

As such, the numerals game activity is designed to tax both math ability as well as all three aspects of executive function: *working memory* (children remember their role and act it out), *inhibitory control* (children suppress the impulse to prematurely switch roles), and *cognitive flexibility* (children flexibly switch roles when they have completed the set of cards).

Tools activity #4: Freeze dance

Some other activities lack an academic component but instead serve exclusively as "attention-focusing" activities (Leong & Bodrova, 2011, p. 13). For example, the freeze dance game prompts teachers to hold up a picture card with a certain body position depicted on it (e.g., hands crossed over one's chest). The children view the picture while dancing;

when the music stops, the teacher obscures the card, and the children strike the pose that had been illustrated on the card.

Thus, freeze dance is designed to target all three aspects of executive function because children have to 1) recall and imitate the card's pose after it has been occluded from view, 2) flexibly switch poses across each card, and 3) inhibit the impulse to strike the pose when they first see the card and instead wait until the music stops.

2.4.2 Summary of Tools program

Vygotsky's influence on Tools is made manifest through each Tools activity. All activities include pre-specified mediators in the curriculum training manual (Leong & Bodrova, 2011); each activity targets children's ZPD either through direct work with the teacher or peers; and, finally, Tools' central focus on play reflects Vygotsky's assertion (Vygotsky, 1933b) that "children achieve their greatest self-control in play" (p. 13).

Thus, overall, all Tools activities are designed to improve children's self-regulation skills. However, this hypothesis has never before been rigorously tested. The aim of this thesis is to do just that, but first, the theoretical basis of this dissertation's analytic approach is described in the upcoming Chapter Three.

CHAPTER 3: Bronfenbrenner's bioecological model

As discussed in section 1.5, this dissertation contains two theoretical paradigms: 1) Vygotsky's cultural-historical theory, which underpins the Tools program, and 2) Bronfenbrenner's bioecological theory, which underpins this dissertation's analyses. Whereas Chapter Two outlined Vygotsky's cultural-historical theory, this chapter details the bioecological framework.

First, this chapter briefly discusses alternative theoretical accounts regarding the roots of self-regulation development. Next, this chapter presents Bronfenbrenner's model and explain its merits in representing self-regulation development within the present context. Finally, this chapter explains how Bronfenbrenner's model informs this dissertation's analytic approach.

3.1 Other theoretical accounts of self-regulation development

Before presenting Bronfenbrenner's bioecological model, it is important to discuss various other theorists who have conjectured about the roots of self-regulation. For example, Freud (1899) wrote about the conflict between a person's impulses (the id) and societal conventions (the superego). The mediator, called the ego, is considered among the first self-regulatory mechanisms posited in the literature (Bronson, 2000). What, then, are the roots of the ego? What can enhance its development? What can undermine its development?

These questions were subsequently addressed by behavioral theorists such as Skinner (1953), who posited that self-regulation, like all behavioral phenomena, develops through stimulus-response relationships. For example, if a teacher punishes a misbehaving child, then the child will stop the undesired behavior. A teacher can then reward the child for exhibiting good self-regulatory behaviors, which will theoretically lead to the subsequent replication of such behaviors.

Thus, the behaviorist account asserts that effective implementation of punishments and rewards leads the student to elicit desirable self-regulatory behaviors. On the contrary, meta-analytic evidence has shown that rewards (Deci, Koestner, & Ryan, 1999) and punishments

(Gershoff, 2002) may induce immediate compliance, but the desired behaviors diminish over time as the incentives fall away. This evidence, which will be expanded upon in the literature review section (Section II), suggests that the behaviorist approach does not result in autonomous self-regulation but rather regulation by external incentives.

In response to the shortcomings of the behaviorist approach, social learning theorists such as Bandura (1977) asserted that children can learn to self-regulate through observation of punishments and rewards without experiencing them firsthand. Specifically, Bandura argued that children's behavior, cognition, and environment reciprocally influence one another.

Bandura (1991) applied this line of thinking specifically to self-regulation, arguing that a child's observations and thoughts affect their self-regulatory behaviors, and the outcomes of those behaviors reciprocally influence their thoughts. Bandura called this inter-relationship the reciprocal determinism model, which is depicted below in figure 4.

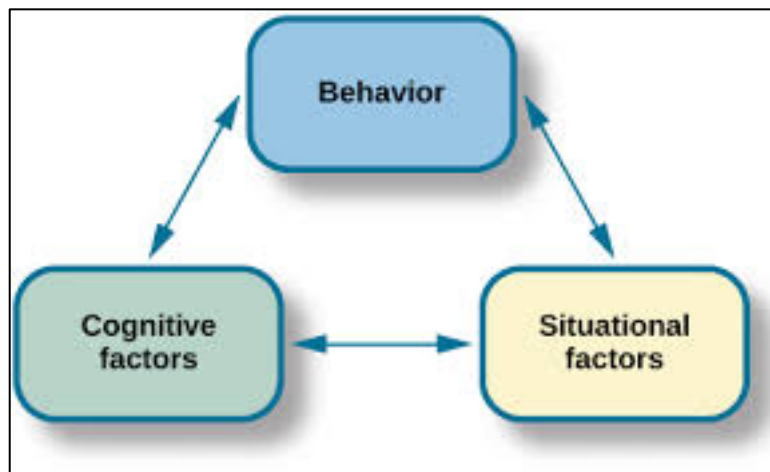


Figure 4: Bandura's reciprocal determinism model

In an illustration of reciprocal determinism called Bandura's Box (1977), Bandura describes a child who does not like school (i.e., a cognitive factor in the model depicted above), which causes the child to misbehave in school (i.e., a behavior). The child's teachers and administrators respond by restricting the child's access to certain elements of school (i.e., situational factors), which then causes the child to enjoy school even less, which subsequently exacerbates the child's behavior issues. For the purposes of this study, the

Tools program could be considered as a situational, or environmental, factor that aims to improve children's cognitive control, which should then reciprocally improve their behavior.

This notion of reciprocal influence is also captured in the transactional model (Sameroff, 1975, 2009) as well as Bronfenbrenner's ecological model (1979). Sameroff's transactional model (1975) posits that children become agents in the social construction of the people and relationships around them. For example, a parent may be predisposed toward harsh parenting styles; however, if the child has a naturally agreeable disposition, then the child's positive behavior could counteract his or her parent's predisposition toward harsh parenting to yield more sensitive parenting. The child and parent thus reciprocally influence the trajectories of one another – neither trajectory is wholly dependent on the other, but both are substantially influenced by one another (Sameroff, 1975).

Whereas Sameroff's transactional model describes two forces that bidirectionally influence one another and Bandura's reciprocal determinism involves three, Bronfenbrenner's ecological model arguably includes a more extensive set of contextual influences on children's development. The full set of contextual influences included in Bronfenbrenner's model is described in section 3.2 below.

In addition to the broader range of contextual influences, Bronfenbrenner's model also accounts for the fact that children are nested within concentric levels of influence, whereas neither Sameroff nor Bandura mention the nested nature of child development. Specifically, Bronfenbrenner theorized that children are nested within and affected by micro-level systems such as family and peers, which are then nested within more macro-level systems such as schools, communities, and society. This nested nature of child development was first captured by Bronfenbrenner's (1979) ecological model (as distinguished from his bioecological model), which is described below.

3.2 Bronfenbrenner's ecological model

Before publishing his ecological model, Bronfenbrenner (1974) criticized his contemporaries for researching "strange behavior of children in strange situations for the briefest possible period of time" (p. 2). Specifically, he argued that much developmental research lacked

ecological validity because children's development was studied in laboratory settings that neglected contextual influences on children's behavior. In an attempt to remedy this issue, Bronfenbrenner aimed to capture those contextual influences through his ecological model (Bronfenbrenner, 1979).

In his original model, which was explicated in *The ecology of human development* (Bronfenbrenner, 1979), the child is said to be situated at the center of a series of concentric rings of environmental influence. Each ring affects the child's development, and each ring also reciprocally interacts with the others (see figure 5).

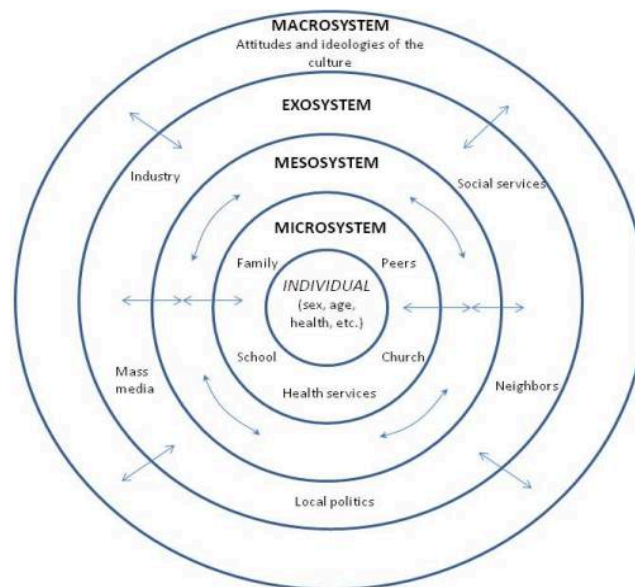


Figure 5: Bronfenbrenner's ecological model

As figure 5 illustrates, the child's first external influences exist in the microsystem, which is comprised of people and structures with which the child has direct contact (e.g., family, schools, and other community organizations). The next layer, the mesosystem, consists of connections among members of the microsystem; essentially, the mesosystem is a "system of two or more microsystems" (Bronfenbrenner & Morris, 2006, p. 817). For example, the mesosystem could manifest as the connection between the child's parents and teachers. Their interaction with one another affects the child's life in school, and the child's life in school reciprocally affects the relationship between parents and teachers.

The fourth ring, depicted by the exosystem, refers to larger structures with which the child does not directly interact but nonetheless remain important for the child's life. For example, the exosystem could refer to structures such as the school district or the parents' workplace. Although the child does not visit the school district office, district policies do directly affect the child's development.

Finally, the fifth ring, called the macrosystem, encapsulates the over-arching socio-cultural values that permeate all levels of the ecological model. Bronfenbrenner argued (1979) that the macrosystem affects children's development as much or more than any other system because all members of the society are shaped by socio-cultural attitudes.

Thus, Bronfenbrenner's original ecological model (1979) included only the environmental influences on child development; the children's personal and biological changes over time were excluded from the model. The omission of the child's personal and biological characteristics manifested as a critical weakness of the theory that Bronfenbrenner himself eventually acknowledged:

Existing developmental studies subscribing to an ecological model have provided far more knowledge about the nature of developmentally relevant environments, near and far, than about the characteristics of developing individuals [...]. The criticism I just made also applies to my own writings [...]. Nowhere in the 1979 monograph, nor elsewhere until today, does one find a parallel set of structures for conceptualizing the characteristics of the developing person. (Bronfenbrenner, 1989, p. 188)

Thus, Bronfenbrenner altered his model to incorporate changes in the child's personal characteristics and biology over time. According to Bronfenbrenner, the model's transformation occurred over an extended period of time, with the new model surfacing in 1986 (Bronfenbrenner & Morris, 2006, p. 794). It is this modified model, called the bioecological model, that was used as the analytic framework for this dissertation.

3.2.1 Bronfenbrenner's bioecological model

After initially criticizing his contemporaries for researching "strange behavior" in "strange contexts" (Bronfenbrenner, 1974, p. 2), Bronfenbrenner's subsequent criticism manifested as the reverse of his original one: "In place of too much research on development 'out of context,' we now have a surfeit of studies on 'context without development'" (Bronfenbrenner, 1986, p. 286). In light of this new critique, Bronfenbrenner created the

bioecological model to include both contextual and developmental influences (Bronfenbrenner & Ceci, 1994).

Although the bioecological model represents a modification of the ecological model, Bronfenbrenner makes clear that much of the former model remains intact:

The present formulation makes no claim as a paradigm shift (if there be such a phenomenon); rather, it continues a marked shift in the center of gravity of the model, in which features of earlier versions are first called into question but then recombined, along with new elements, into a more complex and more dynamic structure [. . .]. In the bioecological model, development is defined as the phenomenon of continuity and change in the biopsychological characteristics of human beings, both as individuals and as groups. (Bronfenbrenner & Morris, 2006, p. 793-794)

Thus, the bioecological model incorporates children's biological and psychological characteristics alongside their contextual influences. To do so, Bronfenbrenner augmented his model with the chronosystem, which signifies the child's biopsychological changes over time and how those changes affect subsequent development (Bronfenbrenner & Ceci, 1994).

Moreover, he incorporated children's biological characteristics into the central ring, which represents the child (Bronfenbrenner & Morris, 2006, p. 794). Thus, the bioecological model closely resembles the original ecological model in form; the bioecological model simply integrates children's biopsychological characteristics, thereby "further differentiating, expanding, and integrating the original 1979 conceptualization of the environment" (p. 796).

Overall, the bioecological model captures a child's characteristics, the diverse influences on the child's development, and the nested nature of the child within those influences. Because self-regulation is a biologically-mediated capacity that develops over time (Diamond, 2006; Heatherton & Wagner, 2012), it is necessary to account for children's biological characteristics as well as their environmental influences. Thus, I have adapted the bioecological model (see figure 6) to capture the research foci of this dissertation.

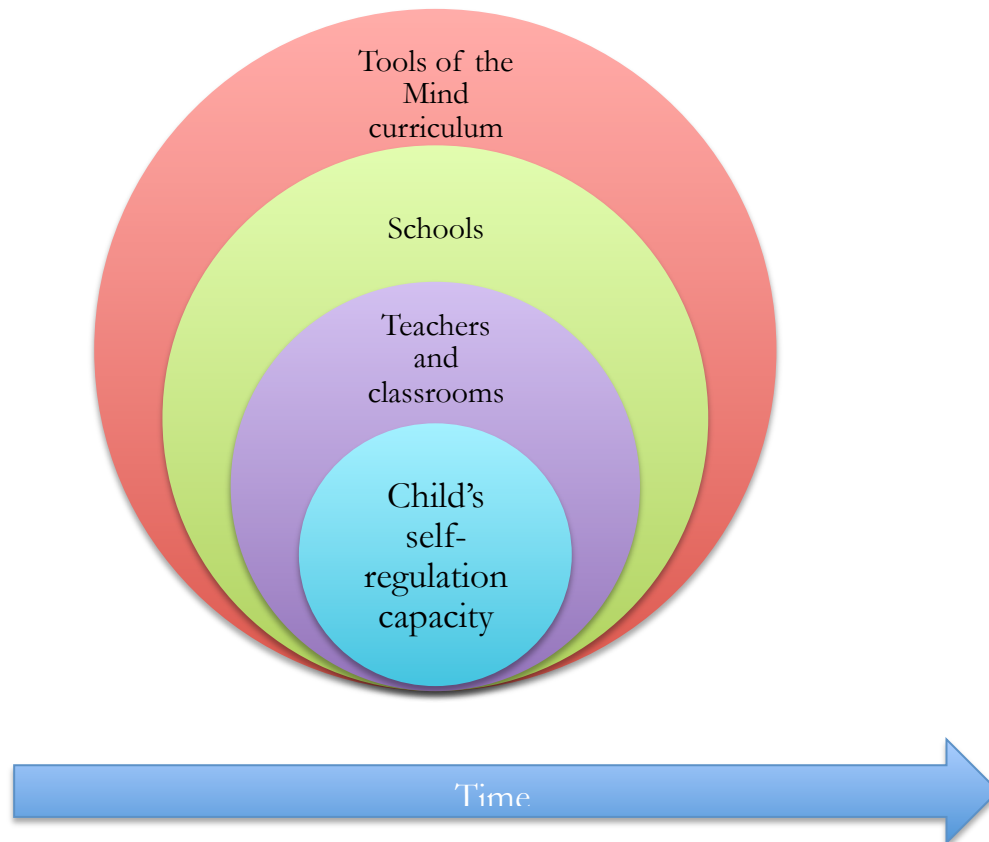


Figure 6: Adapted bioecological model for this dissertation

Figure 6 illustrates how children's self-regulation (the central circle) is affected by the educational factors surrounding them. In the microsystem (blue section), teachers and children affect one another's instructional and self-regulatory behaviors, respectively. Next, the exosystem (green section) represents the interactions between the school and the child. After that, the Tools curriculum represents the macrosystem (red section), which serves as the over-arching Vygotskian philosophy that affects all the lower-level rings. Finally, the blue arrow underneath the rings represents the chronosystem, which captures children's self-regulation changes over time.

In the original bioecological model, Bronfenbrenner argued that interactions among levels are equally important to the subjects within the levels (Bronfenbrenner, 1979). Or, as Bronfenbrenner phrased it: "the main effect is in the interaction" (p. 3). As with Bronfenbrenner's bioecological theory, all the rings in figure 6 collectively influence the child's development and one another.

For example, Tools influences the school environment as well as teacher practices and children's self-regulation; the teacher's instructional practices influence children's self-regulation development; and, at the center, the child's self-regulation both receives influence from all outside systems while also reciprocally influencing them.³

It is important to note that the adapted bioecological model depicted in figure 6 omits elements such as parental influence, community composition, and other societal factors. This is not because those factors do not influence children's self-regulation but rather that the data available in the present study do not allow for such analyses. The characteristics of this dissertation's data will be described in the methodology sections (Chapters Eight and Eleven for study one and study two, respectively). At this point, section 3.2.2 below briefly explains the bioecological model's application to this dissertation's analyses.

3.2.2 Bioecological model's application to the present analyses

Before describing how Bronfenbrenner's model underpins this dissertation's analyses, it is first worth noting that Bronfenbrenner's views regarding development in context were substantially influenced by Vygotsky (Bronfenbrenner, 1986). As described in Chapter Two, Vygotsky was among the first theorists to emphasize the role of context on development, which is a critical focus of Bronfenbrenner's model as well.

Beyond the ideological overlap between Bronfenbrenner and Vygotsky, though, the most important reason to incorporate Bronfenbrenner's bioecological model into this thesis is that the model aligns closely with my analytic method (see Chapters Eight and Eleven). Essentially, this dissertation will analyze children nested within classrooms nested within schools, all of which are nested within the Tools program. Moreover, this dissertation involves longitudinal self-regulation data, which can be captured in the chronosystem.

³ Although Bronfenbrenner created the model, John Dewey's words also accurately capture the model's essence: "Now the change which is coming into our education is the shifting of the center of gravity. In this case, the child becomes the sun about which the appliances of education revolve; he is the center about which they are organized" (Dewey, 1915, p. 103).

Thus, the data to be used in this dissertation largely conform to the multilevel, nested structure of the bioecological model, which illustrates the Bronfenbrenner model's applicability to this dissertation's analytic approach. Moreover, because children's cognitive self-regulation skills are mediated by genetics and neurobiology (see Chapter Five for more information), Bronfenbrenner's bioecological paradigm can effectively model how a child's biopsychological characteristics such as self-regulation interact with his or her nested contextual influences.

This interplay between self-regulation and contextual influences, especially education, will be the focus of this dissertation, which illustrates why the bioecological model is appropriate to underpin the upcoming analyses. Before sharing the analyses, however, we first encounter the upcoming literature review (Section II), which presents existing research regarding self-regulation as well as the contextual factors that influence its development.

SECTION II: LITERATURE REVIEW

CHAPTER 4: Why is self-regulation important?

This chapter reviews the literature regarding self-regulation during childhood and throughout the life course. As the title of this dissertation suggests, I aim to investigate whether and how Tools affects children's self-regulation development. Why, then, is it important for children to develop autonomous control over their attention and behavior? In other words, why is self-regulation so critical? The following sections detail the evidence pertaining to various correlates of self-regulation inside the classroom and beyond.

4.1 Self-regulation and academic achievement

Among the first and most famous tests of self-regulation was Mischel et al.'s (1972) marshmallow study. The task required preschool children to sit with a marshmallow for 15 minutes without eating it. If they succeeded, then they received a second marshmallow. Thus, children had to inhibit a dominant impulse (i.e., to eat the marshmallow) and instead follow a subdominant impulse – delay gratification and earn another marshmallow (Mischel et al., 1972).

Years later, follow-up tests indicated that children who had waited the full 15 minutes without eating the marshmallow scored, on average, 210 points higher on America's Scholastic Assessment Test (SAT) than those who had been unable to wait (Cherniss, 2000). Research on that first sample of 'marshmallow test' children was among the first to suggest the connection between self-regulation and favorable academic outcomes.

But why should this be the case? Why would children's ability to inhibit impulses predict higher achievement? As the social interaction theorist Bandura (1977) wrote, "self-regulation directs learning processes from beginning to end. From the selection of stimuli to attend through the monitoring of attention and retention through decisions to retrieve knowledge or reproduce skills, self-regulation produces an executive function" (p. 13).

As Bandura (1977) wrote, it is executive function, or cognitive control, that drives learning. If students can control their thinking, then they can selectively attend to important information, ignore distractions, exercise logical memory, and utilize other skills that

promote academic achievement across multiple academic domains (Blackwell, Cepeda, & Munakata, 2009; Clements, Sarama, & Germeroth, 2016; Gathercole, Pickering, Knight, & Stegmann, 2004). The section below further highlights the connection between executive function and academic skills such as literacy and numeracy.

4.1.1 Connection between executive function with literacy and numeracy

Early literacy and numeracy both require children to flexibly switch between rule sets and inhibit one symbolic representation in favor of another depending on context. For example, the phonics tradition in literacy teaches children the sounds of both ‘t’ and ‘h’ separately; subsequently, though, children learn the ‘th’ sound, which is different than either the ‘t’ and ‘h’ sounds by themselves. Thus, children must use executive function to inhibit the impulse to say the ‘t’ sound when they see the /t/ and instead notice the adjacent /h/, which requires them to vocalize a different sound.

Executive function is similarly critical for math ability (Bull & Scerif, 2010; Gilmore, Attridge, Simms, & Inglis, 2013). For example, children facing a numerical operation with two numbers must learn to manipulate those numbers differently depending on the sign that divides them (e.g., a minus versus a plus sign). This experience taxes all three elements of executive function: 1) working memory to remember the relevant operation and execute it, 2) cognitive flexibility to switch among the various operations, and 3) inhibitory control to suppress the impulse to perform one operation over another.

Similarly, mathematical exercises in ordinality, cardinality, transitivity, and pattern completion require robust executive function skills (Blair & Raver, 2015). For example, with pattern completion, children must remember relevant information (e.g., recalling what previous parts of the pattern look like), switch attention across rule sets (e.g., identifying new guiding principles across different patterns), and inhibit distracting information (e.g., ignoring misleading parts of a pattern to focus on the guiding principle of the pattern).

As Blair and Raver (2015) assert, an interest in self-regulation “does not supplant interest in the development of acquired ability, such as early knowledge of letters and numbers; it sets the stage for it” (p. 711). In other words, self-regulation skills do not take the place of

academic skill development. Rather, by promoting children's ability to control their mental processes, educators can "set the stage" (p. 711) for children to excel academically.

It is now more intuitively clear how executive function would be associated with higher academic achievement. Many learning tasks require robust working memory, cognitive flexibility, and inhibitory control skills for optimal performance. Despite this intuitive link between executive function and academic skills, does the empirical evidence actually align with the theoretical intuition?

Indeed it does. Whereas Mischel et al.'s (1972) marshmallow study children exhibited higher academic achievement years later on the SAT, other research (Duckworth & Carlson, 2013; Fitzpatrick, McKinnon, Blair, & Willoughby, 2014; Ponitz, McClelland, Matthews, & Morrison, 2009; Schmitt, McClelland, Tominey, & Acock, 2015) indicates immediate academic benefits for young children with robust self-regulation.

Even during early childhood, children with high self-regulation have been shown to attain higher preschool achievement scores (Barnett et al., 2008). One study (Ponitz et al., 2009) measured students' self-regulation using an executive function task called "Heads-Toes-Knees-Shoulders" (HTKS), which requires children to flexibly switch between different sets of rules and inhibit dominant impulses.⁴ The authors found that higher HTKS pre-test ratings in autumn predicted significantly higher spring literacy ($\beta = .87, p < .01$) and math ($\beta = .79, p < .01$) scores in their sample of 343 American kindergarten children.

Despite the associations between children's self-regulation and academic achievement, some may argue that researchers have simply confounded executive function with IQ. That is, the observed gains for children with high executive function in the Ponitz et al. (2009) study could be a product of those children's IQ, for which Ponitz et al. did not control. Thus,

⁴ Heads-Toes-Knees-Shoulders involves touching the correct body part based on the teacher's instructions, which change after each round. Specifically, children touch a body part opposite to the one articulated by the examiner (e.g., if the teacher says 'heads,' then children should touch their toes). This activity engages all aspects of executive function: 1) working memory (remembering the teacher's directions and acting upon them), 2) cognitive flexibility (switching among the rules as they change during each round), and 3) inhibitory control (not touching the body part that you hear, but rather the body part that the teacher has previously specified through a rule).

section 4.1.2 below analyzes the associations between self-regulation and IQ to determine the relative contribution of each toward academic achievement.

4.1.2 Self-regulation versus IQ

It may be that self-regulation predicts academic achievement, but it is also important to consider its predictive power vis-à-vis other capacities. For example, multiple studies show a strong association between IQ and academic achievement (Calvin, Deary, Webbink, Smith, & Visscher, 2012; Kaufman, Reynolds, Liu, Kaufman, & McGrew, 2012). So how do children's self-regulation skills compare with IQ when predicting academic outcomes?

A seminal paper by Duckworth (2005) investigated that precise question. With two consecutive samples of 140 and 164 eighth grade students, respectively, Duckworth measured students' self-regulation using self-reports, parent reports, and teacher reports. Students also filled out monetary choice questionnaires to assess their gratification delay and metacognitive abilities. With both student cohorts, the authors analyzed whether self-regulation or IQ accounted for more variance in students' final grades, school attendance, standardized assessment scores, and selection into a competitive high school program (Duckworth & Seligman, 2005).

The authors found that self-regulation accounted for more than twice as much variance in students' final grades than did IQ ($\beta = .65, p < .001$ versus $\beta = .25, p < .001$, respectively). The relative predictive superiority of self-regulation remained constant across all outcomes (Duckworth & Seligman, 2005); specifically, students with robust self-regulation also had significantly higher school attendance, standardized achievement test scores, and rates of competitive high school entry than did those with lower self-regulation (p. 940).

Another study (Wolfe & Johnson, 1995) compared the role of children's measured self-regulation versus their SAT scores to predict students' university grade point average. In addition to self-regulation, the authors examined 31 other "personality variables" (Wolfe & Johnson, 1995, p. 177) such as extraversion, self-esteem, energy, and others to determine which variables predicted university grade point average.

The authors found that self-regulation explained 9% of the variance in students' grade point average compared to 5% by SAT scores (R^2 difference was significant at $p < .05$). Similar to Duckworth's (2005) study, self-regulation accounted for nearly double the variance of the standardized test scores. Moreover, in the analysis of the 32 personality variables, self-regulation emerged as the only one to significantly predict university grade point average. Thus, not only is self-regulation a relatively strong predictor vis-à-vis IQ and standardized achievement measures, but it is also a strong predictor compared to other socio-emotional skills (Heckman & Masterov, 2007; Wolfe & Johnson, 1995).

4.1.3 Longitudinal evidence

While the aforementioned two studies (Duckworth & Seligman, 2005; Wolfe & Johnson, 1995) identified self-regulation as a strong predictor of academic achievement, both studies had cross-sectional designs. In order to establish that self-regulation abilities predict improved academic achievement over time, longitudinal study designs are necessary.

Longitudinal modeling enables researchers to establish directionality (i.e., that self-regulation is predicting higher academic achievement as opposed to the converse) because prior levels of each capacity can be held constant while predicting the other (Kenny, 1975; Punch, 2014). That is, longitudinal research can control for initial levels of self-regulation to determine whether self-regulatory gains are associated with subsequent academic achievement gains, and vice versa.

The self-regulation field has abundant longitudinal research to illuminate the benefits of self-regulatory capacity (Moffitt et al., 2011; Neuenschwander, Röthlisberger, Cimeli, & Roebbers, 2012; Ning & Downing, 2010; Nota et al., 2004); the one with the most generalizability and power is arguably Duncan et al. (2007), who analyzed six longitudinal datasets – two nationally representative datasets from the US, two from multi-site studies of American children, one British dataset, and one Canadian dataset. The study's aim was to examine the relative contributions of attentional control, social-emotional skills, and academic skills at school entry on subsequent math and literacy achievement.

Beyond early literacy and mathematics skills, which are consistently the strongest predictors of subsequent literacy and math skills (Bull & Scerif, 2010; Duckworth & Seligman, 2005), attentional control emerged as the next strongest predictor of future academic achievement ($\beta = .07, p < .05$). That relationship remained significant even after the authors controlled for socio-economic status, gender, and family structure (G. Duncan et al., 2007).

The authors concluded that the underlying reason behind the effect may be that attention skills “increase the time children are engaged and participating in academic endeavors and learning activities” (G. Duncan et al., 2007, p. 8). That quote is reminiscent of Blair and Raver’s (2015) idea that self-regulation “sets the stage” (p. 711) for academic achievement. That is, with increased attentional control, children can maximize the time that they ignore distractions, remain on task, avoid inhibition errors in math and literacy operations, and generally excel in school.

Ultimately, Duncan et al. (2007) state that attentional control is critical to subsequent academic achievement, but they acknowledge that their results “say nothing about the types of curricula that would be most effective in promoting these skills” (p. 20). This dissertation aims to address that very question.

4.2 Compounding benefits of self-regulation throughout the life course

In addition to academic gains, the benefits of self-regulation have been shown (Flouri, Midouhas, & Joshi, 2014; Moffitt et al., 2011; Sitzmann & Ely, 2011) to extend into people’s behavioral, professional, and financial futures. For example, self-regulation has been shown (Flouri et al., 2014; Raver et al., 2013) to attenuate the link between poverty and externalizing behaviors (e.g., anger, fighting). That is, whereas impoverished children are more likely to exhibit externalizing behaviors, low-income children with high self-regulation have externalizing rates as low as their affluent counterparts (Flouri et al., 2014).

Thus, children who are able to regulate their behavioral and emotional states can decrease behavioral problems that threaten performance in school, and, eventually, in society (Ursache, Blair, & Raver, 2012). For example, one longitudinal study (Moffitt et al., 2011) followed 1000 people from birth to age 32 and found that self-regulation was positively

associated ($p < .05$) with health and financial outcomes while negatively associated with criminal behavior and substance abuse. The analysis controlled for intelligence and social class, thus isolating the contribution of self-regulation on these life outcomes.

Moreover, other studies have indicated that strong self-regulation skills can shrink income gaps between rich and poor (Evans, 2003), increase employment prospects (Daly et al., 2015), improve body mass index indicators (Schlam et al., 2013), and even promote marital stability (Mischel, 2014). Thus, the benefits of self-regulation extend far beyond the classroom and into multitudinous dimensions of life.

4.2.1 Summary

The research summarized above includes both cross-sectional and longitudinal evidence on the importance of self-regulation. Although the research base on self-regulation's benefits is far more expansive than the studies reviewed here, this chapter is intentionally brief. That is because this dissertation does not seek to add to the literature that shows how self-regulation can positively affect people; rather, the rest of this literature review, and the rest of this dissertation, seeks to explore the converse question – how can people positively affect their self-regulation development?

Specifically, this dissertation investigates the Tools curriculum and its impact on children's self-regulation. Before proceeding to the review of existing early childhood interventions and curricula, though, the upcoming Chapter Five first explores other biological and contextual factors that influence children's self-regulation development.

CHAPTER 5: Biological and contextual factors that influence self-regulation development

Although this dissertation is centrally concerned with 'Tools' influence on self-regulation, Bronfenbrenner's bioecological model (see Chapter Three) illustrates the reciprocal interplay among children's biological characteristics and contextual influences (e.g., school, family, community, and society). Thus, this chapter will review the literature regarding various biological and contextual influences on self-regulation.

Specifically, this chapter identifies four factors that influence self-regulation development: 1) genes and neurobiology, 2) socio-economic disadvantage, 3) parenting, and 4) characteristics of children's educational settings. In the following chapters, the role of self-regulatory interventions (Chapter Six) and early childhood curricula (Chapter Seven) are more thoroughly reviewed.

Of course, the set of biological and contextual factors described here does not address all possible factors that influence self-regulation development. As with any developmental phenomenon, the list of potential influences remains unknown and potentially unknowable. Instead, the sections below provide background on four factors commonly discussed in the self-regulation literature as strong predictors of children's self-regulation development.

5.1 Genes and neurobiology of self-regulation

Given that self-regulation (and all cognitive processes) arise from brain functioning, it is useful to understand which neural regions mediate self-regulatory functioning. Thus, sections 5.1.1 through 5.1.3 below outline the genetic bases for self-regulation as well as the neurobiological bases in order to more deeply understand self-regulation's roots.

5.1.1 Genetic bases for self-regulatory capacity

Before a child is born, genetic differences in self-regulation are coded into his or her DNA (Blair & Diamond, 2008). That is, different children's genes code with more or less sensitivity to neural receptors called catecholamines and glucocorticoids, which influence

attentional, emotional, physiological, and behavioral control in children after birth (Tarullo, Obradovic, & Gunnar, 2009). Thus, parents' genetic profiles provide their child with a self-regulatory template upon which further neurobiological development and contextual influences act.

Studies with geographically separated twins (Anokhin, Heath, & Myers, 2004; Friedman et al., 2008; Luciano, Wright, Smith, Geffen, & Martin, 2001) have investigated the level at which children inherit self-regulatory profiles from their parents. Research designs with geographically separated twins are useful because children with identical genetic codes but different environmental circumstances can be tested on the same self-regulation measures. Thus, when genetic codes are held constant across the two twins, the relative contribution of children's genes versus environment on their self-regulation scores can be quantified through statistical analyses.

Although many self-regulation twin studies share similar designs, results have varied considerably across studies. For example, Luciano et al. (2001) found that 43% of the variance in children's working memory skills could be explained by their genes, whereas Anokhin et al. (2004) found that 60% of the variance in children's inhibitory control skills could be explained by their genes. Even more dramatically, a twin study conducted by Friedman et al. (2008) concluded that individual differences in self-regulation "are almost entirely genetic in origin" (p. 201), with over 90% of the variance explained by genes.

Although genetic studies (Friedman et al., 2008; Miyake & Friedman, 2012) suggest that self-regulation is highly heritable, Miyake & Friedman (2012) emphasize that "heritability does not mean immutability" (p. 11). That is, even though children inherit a baseline self-regulatory capacity from parents' genetic code, children's self-regulation can improve or deteriorate based on contextual influences. After birth, contextual influences affect the self-regulatory neurobiological mechanisms, which are now discussed below.

5.1.2 Neural structures underpinning self-regulation

The brain region that most directly mediates self-regulatory functioning is called the prefrontal cortex (PFC), which is situated behind the forehead (Diamond, 2006). The PFC

enables people to engage in executive function – literally, the ability to serve as a master, or ‘an executive,’ over attention, emotions, and behavior (Bell & Deater-Deckard, 2007). When PFC functionality is compromised, whether through alcohol consumption, stress, poverty, or other factors, self-regulation is compromised with it (Duckworth & Steinberg, 2015).

Although the PFC, broadly speaking, controls executive function, the PFC can be subdivided into smaller neural regions that mediate various parts of self-regulation. For example, neuroscience research (Bell & Deater-Deckard, 2007; Diamond, 2006) indicates that the structures underlying emotional and attentional control are in fact distinct. Specifically, within the PFC, the anterior cingulate cortex (ACC) contains both orbitofrontal and dorsolateral cortices. The orbitofrontal region is associated with emotional regulation, whereas the dorsolateral is associated with attentional regulation and cognitive control more generally (Zelazo & Carlson, 2012).

Thus, although the ACC enables both emotional and attentional regulation, they are mediated by slightly different brain structures. Consequently, the two can be treated as slightly different concepts, which is what this dissertation has opted to do through its focus on cognitive self-regulation (as opposed to emotional regulation). This logic also surfaces in Duckworth & Steinberg (2015), who argue for conceptual separability between emotional and attentional regulation based on their distinct neural underpinnings. That is, if different brain structures underlie different self-regulatory skills, then those self-regulatory skills can be considered as conceptually distinct from one another.

Similarly, Bell writes (2007) that the neural structure of self-regulation “allows the conceptualization of self-regulation as a unitary construct with multiple levels” (Bell & Deater-Deckard, 2007, p. 414), including attentional, behavioral, and emotional components. She goes on to explain that developmental researchers typically “focus only on one or two conceptual levels at a time” (p. 409). Thus, this dissertation’s focus on two levels of self-regulation (i.e., attentional and behavioral rather than emotional) comports with the approach in much of the developmental field.

Despite the robust research base underlying the biology of self-regulation, scientists still have much to uncover regarding neurobiological variability across children. That is, biological

research often implicitly assumes that self-regulatory neural processes function similarly across children; however, “this assumption is rarely tested, yet it is just as plausible (and testable) to propose that individual differences in self-regulation stem from heterogeneous processes, operating in distinctive ways for different subgroups of individuals” (Bell & Deater-Deckard, 2007, p. 415).

For example, some children may have more cognitive control than emotional control, whereas others may exhibit the reverse. This would have concrete implications for children; attentional deficits are associated with physiological problems such as high resting heart rate and blood pressure (Biederman, Hirshfeld-Becker, & Rosenbaum, 2001), whereas emotional regulatory deficits are associated with anxiety and aggressiveness (Calkins, S. D., Howse, R. B., & Philippot, 2004). Thus, more research is necessary to demonstrate how neurobiological variation across children manifests as attentional, behavioral, emotional, and physiological issues in later life.

5.1.3 Summary of genetic and neurobiological influence on self-regulation

To summarize, it is important to reiterate Miyake et al.’s (2012) words: “heritability does not mean immutability” (p. 11). When reviewing research about the genetic and neural underpinnings of self-regulation, the sense that cognitive control capacity is biologically fixed can seem logically appealing. That conclusion does not, however, align with the evidence. Neurobiological research from Posner and Rothbart indicate that executive function training can result in changes within brain physiology (Posner & Rothbart, 2007). That is, people’s experiences can, and do, catalyze biological change.

Given self-regulation’s malleability, the next step is to identify experiences and contextual variables that affect self-regulation development. Thus, the following sections detail relevant contextual influences that are associated with both self-regulation growth and decline.

5.2 Contextual factors that affect self-regulation

Once again, the list of contextual influences on self-regulation, or any developmental phenomenon, is potentially endless. Human development can be affected by virtually

anything with which we come in contact, thereby complicating the exhaustive investigation of all contextual factors that influence self-regulation.

Nevertheless, the three sections below sequentially outline the roles of poverty, parenting, and characteristics of educational settings on children's self-regulation. These contextual correlates, as well as the self-regulation interventions and early childhood curricula to be described in Chapters Six and Seven, respectively, are often cited (Blair & Raver, 2015; Duckworth & Carlson, 2013; Karreman et al., 2006; Landry, Miller-Loncar, Karen, & Swank, 2014; Raver et al., 2013) as exerting substantial influence on self-regulation development.

5.2.1 Self-regulation and poverty

Poverty ranks as one of the most consistent contextual predictors of self-regulation decline (Raver et al., 2013; Tominey & McClelland, 2011; Ursache et al., 2012). Specifically, poverty is associated with the “double jeopardy of reduced learning opportunities and reduced support for self-regulation development to foster learning and engagement” (Blair & Raver, 2015, p. 723). That is, children in poverty are less likely to access early education programs, which are expensive in many parts of the world, and are more likely to live in chaotic environments, which can undermine self-regulation development (Tough, 2012).

The ‘chaotic environments’ referenced in the previous paragraph are characterized by economic instability, neighborhood dangers, and other stressors associated with poverty, which all can imperil self-regulation growth (Perry & Szalavitz, 2006; Raver et al., 2013). Indeed, it is not necessarily poverty in and of itself that threatens self-regulation; rather, self-reported levels of psychological distress mediate the relationship between poverty and diminished self-regulation (Duckworth, Kim, & Tsukayama, 2013).

Specifically, high levels of stress can be measured by children's allostatic load, which is a summary index created by McEwen & Stellar (1993) that quantifies how much stress affects one's physiology. When children in poverty repeatedly experience environmental stressors, their allostatic load increases, which research has shown (G. W. Evans, Kim, Ting, Tesher, & Shannis, 2007; Gary W. Evans, 2003) to directly harm their self-regulation growth.

The process involves three steps: 1) children's repeatedly experience stressors associated with poverty, 2) these stressors trigger a neuroendocrine response from children's hypothalamic-pituitary-adrenal (HPA) axis, which is designed to modulate stress responses more sparingly than what typically occurs among children in poverty, and 3) the overly frequent neuroendocrine responses from the HPA axis compromise PFC capacity, which undermines children's self-regulatory functioning (Perry & Szalavitz, 2006). Through this process, the attendant life circumstances associated with poverty collectively threaten children's self-regulation development.

Although poverty compromises self-regulation growth, self-regulation has also been shown (Flouri et al., 2014; Raver et al., 2013) to protect children against the stressors of poverty. As mentioned in Chapter Four, low-income children with high self-regulation have externalizing behavior problem rates as low as their wealthier counterparts (Flouri et al., 2014). Thus, although poverty typically predicts diminished self-regulation, the latter can moderate the effects of the former; put differently, impoverished children with robust self-regulation skills can avoid some of the unfavorable outcomes associated with poverty.

How, though, can impoverished children develop strong self-regulation? Poverty predicts lower self-regulation skills, and yet it is precisely those self-regulation skills that can offset the effects of poverty. In some cases, poor children may be genetically or neurobiologically predisposed toward robust self-regulation skills. In the absence of such fortune, however, contextual influences such as parenting and education emerge as potential protective factors. It is these two influences toward which we now turn.

5.2.2 The effect of parenting on self-regulation

As Denckla (1996) observed, "the difference between the child and the adult resides in the unfolding of executive function" (p. 111). That is, as children become able to control their minds and behaviors through executive function, then they start to resemble the adults they will become.

Similarly, Vygotsky (1994) referred to parents as "the ideal form" and children as the "present form" (p. 366) of cognitive control because children are not yet able to exercise

impulse control and voluntary attention in the way that most adults can. For example, children often fail to inhibit impulses that lead to unfavorable outcomes, such as snatching a toy or getting distracted when crossing the road. In such instances, parents employ various strategies so that the child understands the present mistake and avoids it in the future.

But which parenting strategies will lead the child toward more autonomous self-regulation? Is it better to delineate and then strictly enforce behavioral limits, or is it better to allow the child to navigate the world more independently? And when disciplining a misbehaving child, is it better to do so compassionately so that the child feels secure, or is it better to employ punitive strategies so that the child knows not to make the same mistake again?

The two sections below analyze how harsh versus sensitive parenting styles differentially predict self-regulation development. Of course, parenting styles can be subdivided into myriad classifications. Nevertheless, I have opted to distinguish between harsh and sensitive parenting as relatively broad classifications to avoid conceptual debates regarding various parenting style frameworks. Thus, the harsh and sensitive parenting styles also encompass other oft-cited parenting styles (e.g., authoritarian, permissive, etc.) as described below.

Harsh parenting

Harsh parenting is characterized (Baumrind, Larzelere, & Owens, 2010; Scaramella & Leve, 2004; Shumow, Vandell, & Posner, 1998) by disciplinary tactics that rely on punishment to redirect a child's behavior. In the United States, it is estimated that 65% of parents routinely spank their children, whereas many more shout at or ignore their children (Tough, 2012). Despite the pervasive use of harsh discipline, the question remains: does harsh discipline cultivate self-regulation, or does it further exacerbate children's self-regulatory problems?

Previous cross-sectional research (Gershoff, 2002; Piotrowski, Lapierre, & Linebarger, 2013; Scaramella & Leve, 2004) indicates that harsh discipline can be effective for eliciting desired behavior in the short-term but diminishes the child's internalization of the parent's lesson (e.g., internalize the lesson that bullying is wrong). That is, children's aversion to punishment may prompt their immediate compliance; however, the child does not actually internalize the lesson that might prevent poor behavior in the future.

Evidence from a meta-analytic review (Gershoff, 2002) of 88 studies revealed that children who routinely experienced corporal punishment exhibited higher immediate compliance ($d = 1.12, p < .05$) but internalized moral lessons significantly less ($d = -.33, p < .001$) than their peers who were not physically punished. The notion of immediate compliance versus internalization is highly relevant to the discussion of self-regulation. Parents often want their child to comply with their demands, but parents across cultures hope that their child can flourish independently in their absence (Boyer, 2012). Children who experience harsh parental discipline may appear to autonomously regulate their behavior, but that regulation actually requires the presence of an external force (i.e., the threat of punishment) to function.

By contrast, as the self-regulation definition in this dissertation (section 1.1.1) indicates, genuine self-regulation requires volitional control of attention, behavior, and executive function for the purpose of goal-directed action (Blair & Ursache, 2011). In other words, the self-regulation skills of interest here entail that children are internally regulated rather than externally regulated.

Although it has been established *what* effect harsh discipline has, it is also important to ask *why* such discipline undermines self-regulation. Two main hypotheses emerge. Firstly, harsh disciplinary practices such as shouting or smacking can amplify the salience of the distressing event for a misbehaving child (Grolnick & Ryan, 1989). If a child is already upset about his or her misbehavior, then the additional stress of a harsh parental reaction may increase the child's negative emotionality (Dennis, 2006), which may cyclically intensify the level of harsh parenting practices (Scaramella & Leve, 2004).

Secondly, evidence from previous studies (Dix, 1991; Grolnick & Ryan, 1989) suggests that harsh discipline often represents a lack of behavioral regulation by parents. When parents cannot manage their own actions, this provides a poor regulatory model for children. Conversely, children of parents who effectively modulate their parenting behaviors have been shown to have superior self-regulation skills than children of parents without robust self-regulation (Kochanska, Coy, & Murray, 2001).

Although harsh discipline can compromise children's self-regulation, research indicates that other parental disciplinary strategies exist that encourage self-regulatory development

(Baumrind et al., 2010; Piotrowski et al., 2013; Shumow et al., 1998; Skinner, 1986). One study (Piotrowski et al., 2013) compared authoritative parenting (i.e., the parent encourages independent exploration but still delineates behavioral boundaries) with authoritarian parenting (i.e., the parent harshly restricts children's autonomy) using a representative sample of 1,141 American families. The authors found that self-regulation was significantly positively associated ($r = .15$, $p < .01$) with authoritative parenting and significantly negatively associated ($r = -.23$, $p < .001$) with authoritarian parenting, which replicated a similar study's results from a low-income context (Shumow et al., 1998).

Fortunately, parenting intervention literature suggests (K. Jones et al., 2007; Sanders & Mazzucchelli, 2013) that parents can enhance their authoritative parenting skills to counteract their children's self-regulatory difficulties. Several intervention studies have shown that parenting programs such as the Incredible Years (K. Jones et al., 2007; Webster-Stratton et al., 2009) and the Positive Parenting Program (Bor, Sanders, & Markie-Dadds, 2002; Sanders, 1999) can reduce harsh parenting and cultivate parents' self-control.

By helping parents improve their own self-regulation, the interventions also enhance children's self-regulation by providing a model of self-regulatory growth (Sanders & Mazzucchelli, 2013; Webster-Stratton et al., 2009). Parents who are more responsive to their children's needs than to their own tendencies classify as 'sensitive,' which is the parenting style described in the section below.

Parental sensitivity

Parental sensitivity is defined by a parent's responsiveness to children's age-appropriate needs (Karreman et al., 2006). For example, when a parent accurately perceives and effectively responds to a child's need to be fed, praised, or consoled, the parent can be said to be sensitive (Skinner, 1986). Beginning with Ainsworth's seminal attachment studies (Ainsworth, 1979; Stayton & Ainsworth, 1973), developmental research has typically indicated that parental sensitivity is associated with higher levels of self-regulation (Shumow et al., 1998; Tronick, 1989).

However, the relationship between parental sensitivity and self-regulation is not consistent across time or levels of sensitivity (Feldman, Eidelman, & Rotenberg, 2014). For example,

several studies suggest a negative association between children's age and levels of parental sensitivity, whereby younger children require and receive more sensitive parenting than do older children (Spinrad, Eisenberg, Silva, Hofer, & Smith, 2012; Stright, Neitzel, Sears, & Hoke-Sinex, 2001).

In the same vein, high levels of parental sensitivity have been shown (Kogan & Carter, 1996; Leerkes, Blankson, & Brien, 2010) to be especially crucial for infants' self-regulation. In one study (Leerkes et al., 2010), maternal sensitivity in response to child distress (e.g., crying or screaming) at 24 months was predictive of children's emotional regulation at 36 months. However, maternal sensitivity toward instances of non-distress (e.g., child sitting passively on the floor) was not predictive of self-regulation at 36 months.

In other words, sensitivity was predictive of self-regulation only when the children were struggling with their own emotional regulation but not during times of emotional stability (Leerkes et al., 2010). Thus, even though sensitivity is most important during toddlerhood (Kogan & Carter, 1996; Kopp, 1982), the role of parental sensitivity changes across situations depending on a child's emotional needs.

Another study (Bernier, Carlson, & Whipple, 2010) investigated how different types of maternal sensitivity predicted toddlers' self-regulation development. The authors measured maternal sensitivity when the children were 12 months old and then assessed the children's executive function skills at 24 months. With a sample of 80 American toddlers, the authors found maternal sensitivity that encouraged children's independence was more highly associated with children's executive function skills than was maternal sensitivity without encouragement toward independence.

Other studies (Carlson, 2003; D. Glaser, 2000) have also shown that sensitivity matched with age-appropriate scaffolding strengthens self-regulation development during early childhood. In a study (Evans, Kim, Ting, Tesher, & Shannis, 2007) regarding the effects of poverty on executive function at age three, the authors found that a supportive and authoritative relationship between the toddler and mother predicted improved executive function at age three, which then moderated the effects of poverty into early adolescence.

Although supportive relationships are critically important for self-regulation development, it is important to note that overly sensitive parenting can turn into permissive parenting, whereby the parent rarely enforces rules with the child. Previous studies indicate (Baumrind et al., 2010; Piotrowski et al., 2013) that permissive parenting can result in self-regulatory deficits that are sometimes larger in magnitude than those from harsh parenting. For example, Piotrowski et al. (2013) found permissive parenting to significantly increase ($\beta = .34, p < .001$) children's self-regulation difficulties, whereas authoritative parenting yielded a reduction ($\beta = -.11, p < .01$) in such difficulties.

In the Piotrowski et al. (2013 study), permissive parenting emerged as the strongest predictor of self-regulation in the entire study, with child age as the next strongest predictor ($\beta = -.13, p < .01$). Again, without boundary enforcement in permissive parenting, the child sometimes does not learn when to inhibit impulses versus when to indulge them, which can ultimately be more harmful to the child's self-regulation development than harsh parenting (Baumrind et al., 2010).

Optimal levels of parental sensitivity not only affect children's self-regulation at home but also at school. For example, in a study (Stright et al., 2001) with 52 American families, parents' level of emotional support at home was used as a predictor of children's self-regulation in their second grade classroom. The authors found that parents' emotional support significantly predicted ($R^2 = .21, p < .01$) children's classroom self-regulation, which has also been shown (Duckworth & Carlson, 2013; Ursache et al., 2012) to predict higher levels of academic achievement. This finding has been corroborated in other studies that analyzed the interplay between children's educational experiences and self-regulation (Jones et al., 2007; Webster-Stratton et al., 2009), which is the topic of the section below.

5.2.3 Characteristics of education settings that foster self-regulation growth

The central aim of this dissertation is to identify whether and how Tools promotes self-regulation skills. As described in Chapter One, this aim derives from a gap in the literature regarding educational activities and characteristics that consistently cultivate self-regulation. Study two of this dissertation (Section IV) will address that aim by being the first to individually test each of the Tools activities.

Although the research base regarding specific activities and self-regulation is sparse, several qualities of education settings have been associated with children's self-regulation growth. Whereas Chapters Six and Seven of this dissertation will outline the evidence base behind self-regulation interventions and curricula, respectively, the remainder of this chapter details the evidence base behind various characteristics of education settings that have been associated with self-regulation development.

Specifically, the sections below pertain to three activities and/or characteristics of educational settings that are relevant in the Tools context. The three components below do not constitute an exhaustive set of educational factors that impact children's self-regulation. Instead, I selected these three education factors based on my reading of the Tools' curricular developers' articles (Bodrova & Leong, 2003, 2007, 2008, 2013) about the educational characteristics that might influence self-regulation development. The three educational factors that are sequentially discussed below include:

- Pretend play
- Structured versus less-structured environments
- Teacher-directed versus child-directed environments

Pretend play

Whereas Chapter Two specifically detailed Vygotsky's views of pretend play, this section presents empirical evidence pertaining to pretend play and self-regulation development. Indeed, within and beyond the Vygotskian tradition, pretend play is often regarded (Bodrova & Leong, 2007; Carlson, White, & Davis-Unger, 2014; Piaget, 1969) as an effective tool to cultivate self-regulation in young children. As Piaget wrote (1969), children in pretend play must inhibit their own thoughts, preferences, and plans, while instead acting out a role of someone other than themselves.

The first paper to empirically test the role of play on child outcomes (Smilansky, 1968) included an intervention to increase both the frequency and quality of low-income Israeli children's make-believe play. Following the intervention, the intervention group children engaged more frequently in make-believe play; however, Smilansky (1968) reported no associated improvement in child outcomes.

A follow-up study with American children (Rosen, 1974) also implemented an intervention to increase the frequency and quality of play among low-income children. In contrast to Smilanky (1968), however, Rosen (1974) found that increased play was associated with improved group problem-solving behavior and cooperation. Although Rosen (1974) did not directly assess self-regulation, her results nonetheless represent the first empirical evidence of a positive association between play and child outcomes (Bierman & Torres, 2016).

In 1977, Saltz, Dixon, and Johnson expanded upon those previous two play studies by incorporating one of the only randomized research designs in the existing play literature. In this study, the authors explicitly distinguished between socio-dramatic play and fantasy play. In socio-dramatic play, children act out real-world scenarios from every day life (e.g., the grocery store, hospitals, etc.). In fantasy play, children re-enact make-believe stories such as *The Three Billy Goats Gruff* and *Little Red Riding Hood* (Saltz, Dixon, & Johnson, 1977).

The 54 sampled children were randomly assigned to either socio-dramatic play, fantasy play, fantasy discussion (where children simply discussed the children's stories instead of acting them out), or a control condition that worked in small groups to do activities "typical of preschools" (p. 370) such as finger-painting, cutting, and looking at picture books. The outcomes included an impulse control instrument to measure self-regulation as well as a cognitive development measure.

The authors found that both the socio-dramatic and fantasy play condition children outperformed the other two groups in both cognitive outcomes and impulse control. The fantasy discussion group children had been expected (Saltz et al., 1977) to outperform control group children because they exercised their imagination through fantasy discussions, but this effect was not observed. However, both play enactment conditions exhibited improved cognitive and self-regulatory outcomes, with the effects of the fantasy play condition as "systematically superior to those of the socio-dramatic play condition" (p. 378).

The distinction between fantasy and socio-dramatic play is relevant because the Tools program focuses children exclusively on socio-dramatic play. A more recent trial (Elias, C. L., & Berk, 2002) specifically investigated the association between the socio-dramatic play

characteristic of the Tools program and self-regulation. The authors assessed children's play and self-regulation skills twice during a one-year period.

Through two naturalistic self-regulation measures including behavior during make-believe play cleanup and self-regulation during the class' circle time sessions, the authors found that higher levels of socio-dramatic play in autumn predicted improved behavior during play cleanup periods but not circle time in the spring. The authors conclude that the findings are "consistent with Vygotsky's view of socio-dramatic play" as something that "assists children in managing and directing their own behavior" (p. 231).

Despite the positive evidence detailed above, a recent systematic review entitled "The impact of pretend play on children's development" (Lillard et al., 2013) found inconsistent associations between pretend play and self-regulation. The authors presented findings on multiple outcomes including executive function, emotional regulation, and delay of gratification; Lillard et al. (2013) concluded that the methodological weaknesses across studies, which were mostly observational designs, emerged as a flaw in the literature.

Some of the studies in Lillard et al.'s (2013) review indicated significantly positive associations between play and self-regulation, whereas many indicated null or even significantly negative associations. With most of the studies in the review being correlational and non-experimental in design, Lillard et al. (2013) concluded that evidence of self-regulation promotion through pretend play is "sparse at best" (p. 23).

One criticism of the Lillard et al. (2013) review is that they investigated the connection between play and child outcomes through a strictly causal framework (i.e., whether there is conclusive evidence that play *improves* child outcomes), even though researchers have noted the difficulty of producing clear causal evidence in educational research (Punch, 2014). In one critical response paper, the authors (Weisberg, Hirsh-Pasek, & Golinkoff, 2013) wrote that Lillard et al. (2013) "distorted the lack of strong evidence into an argument for their being no evidence for playful learning – a conclusion we believe to be unwarranted" (p. 36).

Similarly, the Tools developers explicitly responded (Bodrova & Leong, 2013) to Lillard et al.'s (2013) findings by arguing that the results "place professionals in a difficult position

because they already face sharp pressure to replace play with academic activities” (p. 111). Instead of replacing play in early childhood, the Tools developers argued against Lillard et al.’s (2013) results by positing a distinction between mature and immature play (Bodrova & Leong, 2013).

Once again, Bodrova and Leong (2013) describe mature play as involving three elements where children: 1) imagine a scenario, 2) plan for and act out specific roles, and 3) abide by the rules dictated by the nature of their specific role. This notion of mature, or structured, play forms the foundation for the Tools curriculum, and study two of this dissertation will be the first to analyze the association between self-regulation and Tools structured play. While Tools is concerned with structured versus unstructured play scenarios, the more general dichotomy between structured versus less-structured activities frequently arises in the early education literature, and the research base behind that dichotomy is described below.

Structured versus less-structured activities

When identifying educational approaches that hone self-regulation skills, at least two conflicting perspectives exist (Bonawitz et al., 2011). In the former, children learn to regulate their own behaviors through regimented, structured activities (Csibra & Gergely, 2009; Kushnir, Wellman, & Gelman, 2008; Swanson, 1999). Structure thus provides children with a model of how to organize their own learning. In the second model, children learn self-regulation through engagement in exploratory, less-structured activities that compel them to create their own structure and define their own learning aims (Krafft & Berk, 1998; Piaget, 1969; Weisberg, Hirsh-Pasek, & Golinkoff, 2013).

During preschool, children likely engage in both structured and unstructured activities. Structured activities may include times where the teacher explicitly directs students’ attention during a story or math worksheet. Less-structured activities may include free play, recess, and free choice center time, when children can select activities from a host of options. Teachers may want to improve children’s self-regulation but be unaware as to whether using more structured versus less-structured activities will optimize students’ development.

One study (Barker et al., 2014) explicitly connected structured versus less-structured leisure time (i.e., outside the school context) with children’s executive function. Barker et al. (2014)

conducted a battery of executive function tests with 70 children (M age = 6.58) and collected parental reports of their children's leisure activities during a one-week timespan. Specifically, parents provided open-ended descriptions of their children's activity for 30-minute blocks between the hours of 5:30 am and midnight; time spent sleeping or in school was excluded. Parents also completed questionnaires pertaining to the child's typical activity patterns throughout the year. Finally, parents indicated whether the child's activity pattern during the target week was "usual or unusual" on a scale ranging 1-7 (Barker et al., 2014).

The results indicated that less-structured activities were associated with higher executive function scores ($\beta = .71, p < .05$). Conversely, more-structured activities predicted lower executive function scores ($\beta = -.59, p < .07$), though the regression coefficient did not reach significance at the $p < .05$ level. The authors controlled for age, verbal ability, and household income, which have all been associated (Flouri et al., 2014; Kopp, 1982; Raver, 2012; L. Skibbe, Connor, Morrison, & Jewkes, 2013) with children's self-regulation skills.

A limitation of the study is its coding scheme for structured versus less-structured activities. Although the coding scheme was adapted from previous studies (Hofferth & Sandberg, 2001; Meeks & Mauldin, 1990), it has not been formally validated (Barker et al., 2014). The researchers coded less-structured activities as those initiated by the child (i.e., play, reading, and spontaneous practice of a skill) or those including frolicsome outings (i.e., sporting events, parks, library visits, and museums).

Researchers coded structured activities as those led by adults, or those that had inherent structure such as homework, studying, religious services, or non-spontaneous practice of a skill (e.g., a piano lesson). Because museum visits, for example, are not clearly less-structured than piano lessons, the authors acknowledge that the coding scheme is "imprecise, and most likely fails to capture important differences across activities" (Barker et al., 2014).

Despite its limitations, Barker et al. (2014) establishes a platform for subsequent analysis. Specifically, they observed that less-structured leisure outside of school time predicts higher self-regulation, but does this association hold within an educational setting? Or does the

classroom promote self-regulation because students must balance their own desires and impulses with those of their teachers and classmates?

One study (Krafft & Berk, 1998) investigated whether closed- versus open-ended learning activities differentially predict verbal self-regulation among a sample of 59 three- to five-year-old American children. The verbal self-regulation outcome measure was private speech, which refers to a child's self-talk that has no social function but instead serves to guide thinking during a task (Vygotsky, 1962). Private speech has been linked (Diaz & Berk, 2014; Winsler, Manfra, & Diaz, 2007) with higher task performance on difficult problems in early childhood and adolescence.

The authors (Krafft & Berk, 1998) defined closed-ended activities as those with only one correct solution that is pre-determined by the materials (e.g., puzzles or worksheets). By contrast, open-ended activities have no pre-determined goal; instead, the child must both define the goal and work toward its fulfillment. Because open-ended activities theoretically enable children to create a goal within their zone of proximal development, the authors hypothesized that open-ended activities would predict increased private speech.

The results supported the authors' hypotheses regarding a positive correlation between private speech and open-ended activities ($r = .38, p < .01$) and a negative association with closed-ended activities ($r = -.31, p < .05$). The results also indicated significant negative correlations between private speech and transition times ($r = -.40, p < .01$), where the teacher explicitly directs students' movement across the classroom. Overall, the authors found private speech to be most prevalent during make-believe play and least prevalent during structured activities such as literacy worksheets (Krafft & Berk, 1998).

As in the Barker et al. (2014) study, the conceptual distinction between closed- and open-ended activities remained ambiguous in Krafft & Berk (1998). Specifically, the authors defined closed-ended activities as those that had a fixed solution, which resulted in both puzzles and worksheets being placed in the same category. However, other studies (Barker et al., 2014; Hofferth & Sandberg, 2001) place puzzle-work in the less-structured category, which leads to continued conceptual confusion between the two categories.

In addition to studies that have examined the relationship between less-structured activities and self-regulation development in preschool, studies with older children have also found (Hannafin, 2004; Schraw, Crippen, & Hartley, 2006) improved student outcomes for children in less-structured environments. Despite the mostly consistent findings in the literature, the difficulty of classifying (Barker et al., 2014; Meeks & Mauldin, 1990) structured versus less-structured activities problematizes much of the research in this area. Thus, future studies that could surmount this classification problem would substantially contribute to the literature regarding structured versus less-structured environments and self-regulation.

Teacher-directed versus child-directed educational environments

Related to the notion of structured versus less-structured environments is the idea of teacher-directed versus child-directed learning. Although teacher-directed activities and structured environments exhibit conceptual overlap (i.e., teacher-directed activities are often highly structured), the literature bases for the two concepts are largely independent, which is why I have separately addressed them here. Specifically, the structured versus less-structured literature focuses more on environmental characteristics (e.g., the structure of a worksheet), whereas the teacher-directed literature focuses more on the actions of the teacher.

For example, one study (Bonawitz et al., 2011) entitled “The double-edged sword of pedagogy” examines how various levels of teacher-directed instruction affect children’s exploratory learning behaviors. In this study, 85 American preschool children (M age = 58 months) were exposed to a novel toy during a trip to a science museum. The toy, made of plastic pipes and lights, had four hidden functionalities that were designed to elicit interest from the children: a squeaker, a light, a mirror, and a music button (Bonawitz et al., 2011).

In the teacher-directed “pedagogical” condition, the adult experimenter demonstrated the toy’s squeak function to the children but did not indicate the toy’s other three functions. In the baseline condition, the children were simply allowed to play with the toy. In both conditions, children played with the toy for as long as they preferred, and the study stopped when the child verbally declared disinterest in continued play.

The results indicated that children in the pedagogical condition played with the toy for significantly less time (M = 119 seconds) than children in the baseline (M = 205 seconds)

condition ($F_{[1, 81]} = 4.52, p < .05$). Further, children in the baseline condition (i.e., no teacher intervention) discovered significantly more functions of the toy ($M = 2.15$) than children in the pedagogical ($M = 1.72$) condition ($F_{[1, 81]} = 4.58, p < .05$).

The authors concluded that “the balance between direct instruction and discovery learning depends largely on the lesson to be learned” (Bonawitz et al., 2011, p. 329). That is, if the learning goal involves immediate awareness of a toy’s squeak function, then a teacher-directed approach may be best. Overall, though, while direct instruction may lead to more efficient learning of a target concept, such teacher-directed learning is also associated with less exploration and creativity by students.

Although the Bonawitz et al. (2011) study provides evidence regarding teacher- versus child-directed approaches to exploratory play, the study does not yield specific insights regarding the pedagogical approach’s impact on self-regulation. By contrast, one study that did rigorously analyze teacher- versus child-directed activity on children’s self-regulation (Qi, Kaiser, & Milan, 2006) included the two teacher approaches as moderating variables in an analysis of language ability and behavioral regulation. With a sample of 69 low-income American preschool children, the authors used teacher report and observational data to monitor children’s behavior during an unstructured free choice block and a teacher-directed learning block (e.g., literacy and math).

The authors (Qi et al., 2006) found a significant main effect ($F_{[7, 122]} = 25.25, p < .001$) between activity type and externalizing behavior problems. That is, children’s behavioral problems differed significantly across teacher- versus child-directed learning time. Post-hoc analyses suggested the boys with low language ability exhibited significantly more behavior problems in structured, teacher-directed activities ($M = 13.07, SD = 7.08$) than in child-directed activities ($M = 2.82, SD = 2.23$).

Because of Qi et al.’s (2006) small study sample ($n = 69$), the authors were unable to use multilevel modeling techniques to account for their clustered data, where teachers rated multiple children within the same classroom. Thus, dependency within the data existed but was left unaddressed (Qi et al., 2006), which can lead to artificially small standards errors and large type I error rates (Tabachnick & Fidell, 2013). This dissertation’s analyses addressed

the data dependency issue using multilevel modeling techniques, as is described in Chapters Eight and Eleven.

Several other studies have investigated the association between teacher-directed and child-directed activities, but many of the papers lack empirical data. For example, one paper (Schraw et al., 2006) investigated the promotion of self-regulation through child-centered inquiry-based learning, but the paper simply advocates for the approach without providing any empirical evidence. Similarly, books from educational practitioners sometimes advocate a more teacher-centered approach (Lemov, 2010) whereas others advocate a child-centered approach (Bodrova & Leong, 2007; S. Goldstein & Brooks, 2007; Rief, 2007), but the books found during my literature search lack empirical evidence to support the effectiveness of either position.

In sum, the research base described here mostly suggests that child-centered activities promote self-regulation more effectively than teacher-directed activities. This idea is explored further in Chapter Seven, which describes the research base behind several prominent early childhood curricula. At this point, however, it seems that more empirical research would be valuable to clarify the relationship between self-regulation and teacher-versus child-directed activities.

5.3 Chapter summary

Once again, the number of biological and contextual influences on children's self-regulation skills is both unknown and potentially unknowable. This chapter reviewed children's genes, neurobiology, socio-economic circumstances, parenting, and characteristics of their educational environments, which all reciprocally interact with one another as described by Bronfenbrenner's bioecological model.

The interplay among those factors could resemble the following cycle: parents pass on a genetic code to their children that gives rise to a certain neurobiological architecture, which is modified by life circumstances such as poverty. Moreover, the parents' own genes, biology, and environment collectively affect their parenting styles, which then affect the children's self-regulation at home and in school. Finally, parenting behaviors at home

directly influence, and are reciprocally influenced by, the child's experiences at school (Jones et al., 2007; Stright et al., 2001).

It is the contextual influence of education toward which we now fully transition. For example, how do children's experiences in school affect their self-regulatory development? Moreover, how can specific interventions and curricula hone students' self-regulation skills? These questions guide the remainder of this dissertation.

CHAPTER 6: Self-regulation interventions

Given the extensive literature documenting self-regulation's benefits, several educational interventions have been designed to improve children's autonomous control. Many interventions target children with issues such as ADHD (Gulchak, 2008; Harris, Friedlander, & Graham, 2005; Regan & Martin, 2013) and oppositional defiance disorder (ODD; Daunic, Smith, Garvan, Barber, & Becker, 2012; Jones, Daley, Hutchings, Bywater, & Eames, 2007; Webster-Stratton, Reid, & Stoolmiller, 2009). While many interventions address chronic self-regulatory deficits, this chapter focuses on programs that target mainstream students, which is the population targeted by Tools of the Mind (Tools).

Specifically, the following sections assess the effectiveness of four school-based early childhood self-regulation interventions: the Incredible Years Teacher Classroom Management Program, the Chicago School Readiness Project, the Head Start REDI program, and the Head Start PATHS program. All four of these interventions are supplementary programs designed to augment a school's pre-existing curriculum.

Although these interventions do not constitute an exhaustive list of early childhood self-regulation programs, they are highlighted in multiple systematic reviews regarding self-regulation interventions (A. Diamond & Lee, 2011; Jacob & Parkinson, 2015), which explains why I have opted to review their research base here. After reviewing the four supplementary interventions, the findings will be summarized in advance of Chapter Seven, which discusses several comprehensive education curricula and their effects on children's self-regulation.

6.1 Incredible Years Teacher Classroom Management Program

The Incredible Years (IY) program aims to minimize children's behavioral problems and promote their academic, emotional, and behavioral competence through parent and teacher training programs (Webster-Stratton et al., 2009). IY originally targeted only children with oppositional defiance disorder and conduct disorder, but the program has expanded to include mainstream populations as well (Webster-Stratton, Reid, & Hammond, 2004).

While IY includes several parent-training programs, it also has a Teacher Classroom Management program (TCM) that aims to improve teachers' classroom management skills in order to enhance children's academic and behavioral outcomes. Given this dissertation's focus on educational practices that improve self-regulation, the evidence surrounding IY's teacher-oriented TCM program is the focus of this section.

Most IY evaluation studies simultaneously analyze multiple IY modules (e.g., the teacher TCM program as well as the parent training program). By concurrently implementing two or more intervention programs, researchers cannot isolate the unique effect of a single program on child outcomes (Hutchings, Martin-Forbes, Daley, & Williams, 2013). Only two published IY evaluation studies (Hickey, McGilloway, Donnelly, & O'Neill, 2015; Hutchings et al., 2013) exclusively analyze the teacher-oriented IY TCM program, both of which were conducted in European contexts. Each is sequentially described below.

6.1.1 Hutchings et al. (2013) IY study

Hutchings et al. (2013) randomly assigned 12 teachers to either the TCM intervention group or a control group that received no additional support. The children's ages ranged from three to seven years old ($M = 57.5$ months, $SD = 6$ months), and all children attended government-funded primary schools in Wales. In autumn, children's behavior was assessed, and each child was classified as either behaviorally problematic or normal; this binary distinction enabled researchers to determine whether the intervention differentially impacted children with varying behavioral profiles (Hutchings et al., 2013).

The study employed a pre- and post-test design whereby teachers rated children's socio-emotional skills in autumn and spring using the Teacher Strengths and Difficulties Questionnaire (TSDQ). Hutchings et al. (2013) also used the Teacher-Pupil Observation Tool (TPOT), which assesses both children's behavior (e.g., off-task behavior, deviance, and non-compliance) as well as teacher behavior (e.g., teacher demands, positive language, and negative language).

Although the study featured both the TSDQ and TPOT rating tools, Hutchings et al. (2013) lacked objective self-regulation measures, where 'objective' signifies that children's scores are

based on their performance on a task rather than a subjective rating from a teacher, parent, or researcher. Thus, both outcome measures in this study were assessor-based reports from either the teacher (TSDQ) or the research team (TPOT).

The multilevel regression analysis indicated some significant benefits for the IY treatment group over the control group. Specifically, from the TPOT measure, off-task behavior was significantly lower ($d = -.53, p < .05$) among the IY group children, whereas levels of child deviance and non-compliance were equal across groups. The other emergent significant effects were for teachers using less negative language ($d = -.36, p < .05$) and children using less negative language toward the teacher ($d = -.42, p < .05$), but those effects were only significant for children who were initially rated as behaviorally problematic.

Furthermore, the authors concluded that the IY group children rated as behaviorally normal at pre-test exhibited an increase in negative language directed at the teacher as well as decreased compliance (p. 582). Hutchings et al. (2013) explained those findings by saying that teachers may have spent more time with the behaviorally challenged students “because [the teachers] were aware that the observers were observing the target children” (p. 581).

Thus, the intervention did differentially impact children with varying behavioral profiles in that children with the highest behavioral needs gained the most from the program.

However, the mainstream students’ decline produced concern among participating teachers and parents (Hutchings et al., 2013). Issues of mainstream IY children exhibiting behavioral decline relative to IY children with behavioral difficulties has also been documented in other IY studies for the parent-training program (Webster-Stratton et al., 2004) as well as the other IY TCM study (Hickey et al., 2015), which is described below.

6.1.2 Hickey et al. (2015) IY study

The second, and most recent, study (Hickey et al., 2015) to evaluate IY TCM was conducted with 445 children within 22 classrooms in Ireland. The teachers were randomly assigned to IY TCM or a control group that received no additional support. Similar to the Hutchings et al. (2013) study, the 445 children were designated by their teachers as having ‘high,’

‘medium,’ or ‘low’ levels of behavioral problems in order to assess whether IY TCM’s impacts varied across students with different behavioral patterns.

Similar to the Hutchings et al. (2013) study described above, the Hickey et al. (2015) study used the TPOT teacher observation tool to assess whether teachers’ instructional practice significantly improved as a result of the IY TCM program. In addition to the TPOT measure, the researchers used a self-report measure that enabled teachers to reflect on their pedagogical practice. Finally, to assess children’s self-regulation outcomes, researchers used the Strengths and Difficulties Questionnaire (SDQ), which has been employed in several developmental research papers (Flouri, Tzavidis, & Kallis, 2010; K. Jones et al., 2007; Malmberg & Flouri, 2011).

The researchers found no significant effects of the IY TCM program on teachers’ observed instructional practice or children’s overall SDQ scores. Whereas IY TCM teachers reported significantly improved pedagogical practice for themselves in the self-report measure, the TPOT teacher observation tool did not corroborate those findings. Once again, the researchers hypothesized that because IY TCM teachers knew that their practice was expected to improve, it is possible that this expectation positively biased their self-reports (Hickey et al., 2015, p. 15).

Moreover, when the data were disaggregated across children’s ‘high,’ ‘medium,’ and ‘low’ levels of behavioral problems, the researchers found a marginally significant ($d = -.14$, $p = .05$) reduction in high-risk children’s overall behavioral difficulties score. This result mirrors the findings for Hutchings et al. (2013), who also found improved self-regulation for high-risk children and diminished self-regulation for low-risk children. The present Hickey et al. (2015) study found the former but not the latter.

In sum, the IY TCM program predicted higher behavioral regulation scores for children with high behavioral needs in both published studies (Hickey et al., 2015; Hutchings et al., 2013). Given that the outcome assessors were not blind to children’s group assignment in either study, it is possible that researchers’ expectations may have biased their ratings. Moreover, both studies lacked task-based, or ‘objective,’ self-regulation measures, which could have addressed issues of rater subjectivity.

Although several other IY evaluation studies exist (Baker-Henningham, Walker, Powell, & Gardner, 2009; Ford et al., 2012; Webster-Stratton et al., 2004, 2009), they all include the TCM program alongside another IY intervention. This dual implementation design precludes analysis of TCM's unique effect on children's self-regulation. Thus, more evidence is necessary to demonstrate whether IY TCM significantly improves mainstream children's self-regulation skills as opposed to only children with existing behavioral problems.

6.2 Chicago School Readiness Project

The Chicago School Readiness Project (CSRP) is a supplementary educational intervention inspired by the IY Teacher Classroom Management program (Raver et al., 2008) described in the previous section. Similar to IY TCM, CSRP is an intervention directed at preschool teachers to improve their classroom management and classroom climate; in contrast to IY TCM, CSRP targets exclusively low-income preschool children (Raver et al., 2008).

CSRP's methods are based on the assumption that honing teachers' classroom management skills will result in more positive teacher-child and child-child interactions (Cooper-Kahn & Foster, 2013; Sutton, Mudrey-Camino, & Knight, 2009). The resultant atmosphere is hypothesized to accelerate self-regulation growth (Raver et al., 2008).

In the only randomized control trial (RCT) CSRP evaluation, the researchers randomly assigned 90 American teachers to receive either CSRP training or continue 'business-as-usual' practice (Raver et al., 2008), where 'business-as-usual' signifies that the teachers continued the same practice they had used before the study began.

As for teacher training, the treatment teachers received five training sessions of six-hours each to improve their classroom practice. By contrast, teachers in the comparison condition received a part-time teacher's aide so that the treatment and control groups had relatively equal levels of additional support.

To measure child-level outcomes, Raver et al. (2011) used both subjective, teacher-report measures as well as objective, task-based self-regulation measures (i.e., children's scores derived from their performance on self-regulation tasks). The objective self-regulation tests included the Peg Tapping task, which requires children to inhibit a response to match a

researcher's tapping pattern, and the Toy Wait task, where children have to resist the urge to use a toy provided by the researcher (Raver et al., 2011).

The results indicated that CSRP children scored significantly higher ($\beta = .28, p < .05$) on the Peg Tapping executive function task than the control children (Raver et al., 2011). The researchers noted a Cohen's d effect size of .37, which is characterized as a medium-sized effect given Cohen's (1988) guidelines. Other reports using the same data sample as the Raver et al. (2011) paper have underscored CSRP's effectiveness in enhancing children's self-regulation (Jones et al., 2013; Raver et al., 2013). In light of these promising preliminary findings, subsequent replication studies to corroborate Raver et al.'s (2011) findings from the original study sample would strengthen CSRP's evidence base.

6.3 Research-based, Developmentally Informed (REDI) Program

The REDI (Research-based, Developmentally Informed) program is another supplementary early childhood educational intervention that aims to hone children's self-regulation skills. REDI developed as an enriched version of the American Head Start early childhood program that focuses on academic achievement, learning engagement, and behavioral regulation for low-income children (Nix, Bierman, Domitrovich, & Gill, 2013). Similar to the Incredible Years and CSRP, REDI is not a stand-alone curriculum but rather a program that seeks to augment the existing curriculum in a school (K. L. Bierman, Domitrovich, et al., 2008).

In one RCT evaluation study (K. L. Bierman, Domitrovich, et al., 2008), the researchers collected data on 356 American pre-kindergarten students in 44 Head Start classrooms over one year. The 356 students were randomly assigned to either the REDI treatment condition or the 'business-as-usual' curriculum condition. Students in the REDI group learned emotional and behavioral regulation strategies, which teachers modeled during 33 'circle time' sessions. Students practiced techniques such as taking deep breaths, articulating their feelings when upset, and talking through problems with peers (K. L. Bierman, Domitrovich, et al., 2008).

In addition to collecting literacy and socio-emotional outcome data, the authors administered autumn and spring assessments to analyze students' self-regulation development (K. L. Bierman, Nix, & Greenberg, 2008). The executive function test battery consisted of four objective self-regulation measures; after the children completed the test battery, the experimenter rated each child's task focus using an observational assessment form (K. L. Bierman, Nix, et al., 2008).

The authors found no significant intervention effects for any of the objective executive function measures. In contrast, the assessor reports of children's self-regulation were significantly more positive ($\beta = .20$, $p < .05$) in the REDI treatment group as compared with the control group. However, it is possible that the ratings were biased due to a lack of blinding (i.e., the researchers knew whether it was a REDI or control classroom when conducting the self-regulation ratings).

In sum, none of the objective self-regulation measures reflected significant gains for the REDI treatment group, and the reliability of the assessor reports is uncertain. Thus, more research is necessary to determine whether the REDI program significantly improves children's self-regulation skills.

6.4 Promoting Alternative Thinking Strategies (PATHS) Program

The final supplementary educational self-regulation intervention discussed here is the Promoting Alternative Thinking Strategies (PATHS) program (Greenberg, Kusche, & Cook, 1995), which is often implemented alongside the REDI program (Arda, 2012). Similar to REDI, PATHS is designed to augment the existing Head Start curriculum in the United States; however, PATHS specifically targets children's social-emotional learning, while REDI is focused more on learning engagement and academic enrichment (Nix et al., 2013).

In an RCT including both PATHS and REDI, Nix et al. (2013) measured children's reading achievement and rated their social behavior. The authors defined social behavior as children's ability to regulate their emotions and work out problems independently (p. 1006). Although the study included no objective self-regulation measures, the authors identify the improvement of children's self-control as among their key aims (Nix et al., 2013, p. 1013).

Using multilevel models with students nested within classrooms, Nix et al. (2013) found that intervention group children showed significantly higher reading ($\beta = .49, p < .01$), emotional understanding ($\beta = .36, p < .05$), and positive social behavior skills ($\beta = .36, p < .001$). In their discussion, the authors explain that the intervention's effect on children's executive function development may account for the academic and social-emotional gains (p. 1014), but Nix et al. (2013) did not directly measure executive function, thus precluding mediation analyses to test this hypothesis.

Although PATHS often serves as a supplement to REDI, PATHS can also be implemented as a stand-alone intervention. One study (Domitrovich et al., 2007) analyzed PATHS implementation in Head Start classrooms with 246 three- and four-year-old children. Teachers were assigned to either PATHS classrooms or ordinary Head Start centers. Teachers were instructed to implement the PATHS socio-emotional activities outside of daily academic activities such as literacy and math.

Unlike the Nix et al. (2013) study, the Domitrovich et al. (2007) study did employ executive function measures including the Peg Tapping and Night/Day task. Both tasks assess children's inhibitory control – that is, children must suppress a dominant response and replace it with a subdominant response according to the experimenter's instructions.

For example, the Night/Day task contains a dark card with a moon as well as a bright card with a sun. Children must say “day” when they see a dark card and “night” when they see a bright card, which taxes their inhibitory control (Domitrovich et al., 2007). Given PATHS' focus on interpersonal and emotional awareness, children were also tested on their emotional knowledge (e.g., classifying a facial expression as happy, sad, angry, etc.) in addition to the objective executive function measures.

The Analysis of Covariance (ANCOVA) results indicated that the treatment group's emotional knowledge scores significantly improved on three of the four post-tests. However, the fact that the researchers used ANCOVA, instead of multilevel models, means that the results do not account for dependency among students in the same classrooms, which casts doubt on these results' accuracy (Domitrovich et al., 2007). Moreover, despite the less conservative estimation process of the ANCOVA vis-à-vis a multilevel model,

scores on both executive function tests were not significantly higher for the PATHS group children relative to the control group.

Thus, similar to several IY studies and the Domitrovich et al. (2008) REDI study, the intervention group children achieved higher self-regulation ratings from assessors but not higher scores on executive function tasks. This is relevant for two reasons. First, given that the task-based, objective executive function measures used in this study have been extensively validated (Ponitz et al., 2009; Smith-Donald, Raver, Hayes, & Richardson, 2007), it is important for PATHS studies to demonstrate improvement on both the task-based executive function measures as well as the informant-report measures.

Secondly, not only did PATHS students not demonstrate gains on the task-based executive function measures, but the observed improvements on the self-regulation assessor rating measures may also be in question. That is, the raters were not blind to the child's assignment condition (i.e., the teachers and researchers knew whether the children were in the treatment or control group); if the raters expected treatment group children to improve, then this might positively bias their post-test self-regulation ratings for the treatment group. Based on these results, it appears that more research is necessary to determine whether the PATHS program significantly improves children's self-regulation skills.

6.5 Summary of supplementary programs associations with children's self-regulation

Based on the aforementioned evidence, the supplementary programs have varied impacts on children's self-regulation development. Each of the four programs has shown some promising results for the development of children's self-regulation. However, given issues of measurement and design within the four programs, the validity and generalizability of those results remains in question.

For example, in the Incredible Years Teacher Classroom Management Program (IY TCM), the positive intervention effects in the Wales and Ireland studies were specific to children with severe behavioral problems, whereas mainstream children showed evidence of self-regulatory stagnation and even decline during the studies (Baker-Henningham et al., 2009;

Hutchings et al., 2013; Webster-Stratton, Reid, & Stoolmiller, 2008). Moreover, the lack of task-based self-regulation measures complicates the determination of whether the program improves other elements of self-regulation (e.g., attentional control, working memory, cognitive flexibility) in addition to teacher-rated behavioral regulation.

Similar to IY, the Chicago School Readiness Project shows some promise in improving children's self-regulation skills; however, it has only one evaluation study from one city conducted by the program developer, which calls its generalizability into question. By contrast, IY has been implemented in over twenty countries, whereas PATHS and REDI have been widely implemented in the United States but lack well-documented effects.

Finally, it is worth noting that both PATHS and REDI are comparatively low-intensity programs with only 1-2 short lessons per week (K. L. Bierman, Domitrovich, et al., 2008; Nix et al., 2013). Domitrovich et al. (2008), the authors of one REDI study described above, assert that perhaps only intensive and comprehensive curricula have the potential to significantly improve children's self-regulation (p. 1814), which has also been noted in other studies (Barkley et al., 2000; Blair & Raver, 2014). Thus, the upcoming chapter presents research regarding comprehensive early childhood curricula and assesses their impacts on children's self-regulation development.

CHAPTER 7: Background on early childhood curricula

This chapter reviews a set of early childhood curricula in order to provide background regarding curricular alternatives to Tools. Some curricula cover only one learning domain (e.g., math), whereas others are comprehensive, meaning that they cover literacy, math, and a range of other academic and socio-emotional skills (Belfield et al., 2006). Tools is a comprehensive curriculum, so the sections below also pertain only to comprehensive early childhood curricula.

As with the interventions described in Chapter Six, the set of early childhood curricula worldwide is effectively innumerable, so the sections below only describe four of the most common early childhood curricula⁵ from the existing literature: 1) HighScope, 2) Creative Curriculum, 3) Bright Beginnings, and 4) Montessori. Specifically, the sections below investigate the evidence base behind these four early childhood curricula in their promotion of children's self-regulatory skills.

7.1 HighScope early childhood curriculum

In the 1960s, the HighScope curriculum was developed in the United States alongside the federal Head Start program, which sought to expand access to early childhood education for low-income families nationwide (Berrueta-Clement, 1984). Thus, HighScope became, and remains, one of the most widely implemented curricula in the United States and much of Europe; the HighScope website estimates that 20% of Head Start centers use the program ("HighScope: Inspiring educators to inspire children," 2015). The website further explains that HighScope is a "comprehensive curriculum" that includes the following content areas:

- Language, literacy, and communication
- Mathematics
- Creative arts

⁵ These four were specifically identified by the *Handbook of child psychology* (Hyson et al., 2006, p. 35) as among the most common early childhood curricula, and they also repeatedly surfaced in my literature searches. Those two factors explain why I chose to present the evidence on these four curricula in this chapter.

- Science and technology
- Social studies
- Learning approaches for students
- Social and emotional development
- Physical development

The curricular website indicates that active and participatory learning form the basis of the HighScope educational approach. That is, students learn best when interacting directly with materials, other children, teachers, and ideas. Additionally, students engage in a ‘Plan-Do-Review’ process each day to determine what they want to learn and how they will approach that goal. Through planning, students exercise goal-oriented thinking that theoretically guides subsequent behavior.

The students then execute the learning goals throughout the day, during which they are meant to remember their learning goals and inhibit impulses to spontaneously shift to other learning activities. Thus, the Plan-Do-Review process aims to strengthen “initiative and self-reliance” in children (“HighScope: Inspiring educators to inspire children,” 2015).

Despite this focus on improving children’s initiative and self-reliance, which can be viewed as conceptually similar to self-regulation, no studies have specifically analyzed self-regulation development in HighScope classrooms. Nevertheless, the Perry preschool study (Berrueta-Clement, 1984), which featured the HighScope curriculum, ranks among the most cited early childhood education studies. The Perry preschool study randomly assigned 123 low-income three- and four-year-old children to either no preschool or to HighScope in the early 1960s. Data were annually collected on the children between ages 3 and 11. The participants were then followed up at ages 14, 19, 27, and 40 (Nores, Belfield, Barnett, & Schweinhart, 2005).

At the age 40 follow-up, researchers found statistically significant ($p < .05$) differences between the intervention and control groups on high school graduation, employment status, wages, incarceration rates, and divorce rates. All of the differences favored the HighScope treatment group. Whereas the HighScope program cost \$12,000 per year per child, Nobel Laureate economist James Heckman estimated societal savings of \$88,000 per participant

(Heckman, Moon, Pinto, Savelyev, & Yavitz, 2010). Thus, those savings translate into an average \$7.33 societal return for every \$1 invested in the program.

Some might expect that the longitudinal timespan from 1962 until 2005 would produce substantial attrition from the original study sample, but this was not the case. Even at the age 40 follow-up in 2005, researchers collected data on 97% of the original participants who had not passed away. Thus, the risk of attrition bias is low. However, the small scope of the study, which included 65 children in the control group and 58 in the HighScope treatment group, diminishes the generalizability of the findings.

Moreover, the study did not specifically collect self-regulation data, which is the focus of the present study. Although HighScope students in the Perry preschool study exhibited higher graduation rates and lower incarceration, divorce, and drug abuse rates, all of which are correlates of robust self-regulation, the original researchers did not collect any self-regulation data on the students (Berrueta-Clement, 1984). This data omission precludes statements about HighScope's impact on children's self-regulation development, which, to the best of my knowledge, has not been directly analyzed.

7.2 Creative Curriculum

The Creative Curriculum is a comprehensive preschool program that targets four developmental domains: socio-emotional, physical, cognitive, and language development. These four domains are targeted through eleven interest areas including art, blocks, dramatic play, toys and games, library, discovery, sand and water, music and movement, computers, cooking, and outdoors ("Creative Curriculum solutions website," 2015).

Creative Curriculum uses a combination of large-group learning, small-group instruction, and teacher-led instruction; teachers then use an online developmental checklist to conduct ongoing assessments of children's progress in cognitive, physical, linguistic and socio-emotional development (United States Department of Education, 2013b). According to the curriculum's website, the program uses "exploration and discovery as a way of learning," which "enables children to develop confidence, creativity, and lifelong critical thinking skills" ("Creative Curriculum solutions website," 2015).

Although several expository articles and manuals exist for Creative Curriculum (Dodge, Colker, & Heroman, 2001; Teaching Strategies, 2010), only one program evaluation study meets the *What Works Clearinghouse* methodological criteria.⁶ That study (Preschool Curriculum Evaluation Research Consortium, 2008), conducted by the United States Department of Education's Preschool Curriculum Evaluation Research (PCER) Consortium, was conducted as a large-scale randomized evaluation of 14 preschool curricula implemented in geographically diverse sites across the United States. The study involved the randomization of 18 classrooms to either a 'business-as-usual' (i.e., control) condition or the Creative Curriculum.

For all 14 curricula tested in the PCER study, children were assessed on both academic and socio-emotional measures. The socio-emotional measures included the teacher-reported Social Skills Rating Scale (SSRS) social skills scale, SSRS problem behavior scale, and the Preschool Learning Behaviors Scale (PLBS). Both SSRS scales (Fantuzzo, Manz, & McDermott, 1998) as well as the PLBS (McDermott, Leigh, & Perry, 2002) are validated instruments of children's behavioral regulation skills. The children were tested on both the SSRS and PLBS at the beginning and end of pre-kindergarten (i.e., the curricula implementation year) as well as a delayed post-test at the end of kindergarten.

The results indicated no statistically significant differences between the Creative Curriculum and the control group students on any of the three behavioral regulation scales at the pre-kindergarten and kindergarten post-tests (Preschool Curriculum Evaluation Research Consortium, 2008). No task-based self-regulation (e.g., Peg Tapping task, Heads-Toes-Knees-Shoulders) data were collected, so future research should include such measures as well when evaluating the Creative Curriculum's effects on children's self-regulation.

In addition to the null effects for self-regulation, the *What Works Clearinghouse* report that reviewed the 2008 PCER study found no significant effects for children's oral language, print knowledge, phonological processing, or math skills (United States Department of

⁶ The *What Works Clearinghouse* aggregates education evaluation data on specific curricula and interventions in order to issue more authoritative conclusions about a program's efficacy. In order to be included in the *What Works Clearinghouse*, a study must have statistical mechanisms to control for potential confounds (e.g., randomized controlled trial, time series, regression discontinuity). Thus, research evidence from the *What Works Clearinghouse* can be considered to be reliable.

Education, 2013b). Thus, more research is necessary to demonstrate that the Creative Curriculum effectively cultivates children's self-regulation skills.

7.3 Bright Beginnings

Bright Beginnings is a comprehensive curriculum that is partially based on both the HighScope and Creative Curriculum models but has a heightened literacy focus (United States Department of Education, 2013a). The curriculum includes the following nine curricular units:

- Language and literacy
- Mathematics
- Scientific thinking
- Technology
- Physical development
- Creative arts
- Social studies
- Healthful living
- Social and personal development

Similar to the Creative Curriculum, Bright Beginnings organizes the classroom to encourage children's exploration of and interaction with adults, other children, and classroom materials. One relevant difference between Bright Beginnings and Creative Curriculum is the former's implementation of a 'family-school connection link,' which requires caregivers to formally express a commitment to engaging in their child's learning (Preschool Curriculum Evaluation Research Consortium, 2008).

Similar to the research background of Creative Curriculum, only one existing study meets the *What Works Clearinghouse's* methodological criteria. That study is the same national PCER evaluation study (PCER, 2008) conducted by the Institute of Educational Sciences in the United States Department of Education referenced in section 7.2. In that study, seven classrooms were randomly assigned to Bright Beginnings, whereas seven other classrooms

were randomly assigned to the business-as-usual condition, which, again, implemented the same curricula that had been in place prior to the study. The same SSRS and PLBS self-regulation measures were collected for the Bright Beginnings evaluation as were collected in the Creative Curriculum evaluation.

The researchers found no significant differences between Bright Beginnings and control group children's SSRS and PLBS scores at the pre-kindergarten post-test or the kindergarten delayed post-test. In addition to the null self-regulation effects, the *What Works Clearinghouse* team that reviewed the 2008 data found no significant effects for children's oral language, print knowledge, phonological processing, or math skills (United States Department of Education, 2013a).

It is worth noting that for all 14 curricula⁷ tested in the PCER study, not a single program significantly improved children's self-regulation scores vis-à-vis the comparator curricula (Preschool Curriculum Evaluation Research Consortium, 2008). Although I could review each program separately in this chapter, it may not be valuable to go through each of the 14 curricula only to share null results. Instead, section 7.4 below shares research results from the Montessori curriculum, which was not tested in PCER but is widely implemented around the world (Institute of Educational Sciences, 2016).

7.4 Montessori

The Montessori curriculum originates from the educational philosophy of Maria Montessori, who originally applied her method to poor children in a Rome housing project at the turn of the 20th century (Lillard, 2005). The Montessori method involves child-directed learning, where students work in mixed-age classrooms on self-selected academic projects. The teacher serves as a facilitator who supports small groups during their work but does not directly instruct the entire class on a given skill. This approach theoretically hones children's

⁷ The full list of the 14 curricula tested in PCER include Bright Beginnings, two versions of the Creative Curriculum, Curiosity corner, Early childhood express, Doors to discovery, Early literacy and learning model, Language-focused curriculum, Let's begin with the letter people, Literacy express, Pre-K mathematics, Project approach, Project construct, and Ready, set, leap.

self-regulation because the children plan their learning goals and then follow through on them with minimal teacher support (Lillard, 2012).

The elements of children's planning resemble those observed in HighScope's "Plan-Do-Review" process ("HighScope: Inspiring educators to inspire children," 2015) as well as 'Tools' play planning (Bodrova & Leong, 2007); that is, children develop self-regulation by preparing a plan and then following through on it without extensive adult monitoring. Despite the theoretical attractiveness behind the Montessori learning approach, the research community has not produced any randomized Montessori trials. In fact, no Montessori studies meet the *What Works Clearinghouse's* methodological criteria for inclusion (Institute of Educational Sciences, 2016).

Dr. Lillard, the most cited researcher in the Montessori literature base, has noted (2005, 2012) this paucity of rigorous Montessori research and conducted several observational studies about the curriculum. Given that over 90% of Montessori schools in the United States are privately-run schools that require tuition payments from students (Lillard, 2012), it is difficult to conduct randomized control trials on the program. Nonetheless, Lillard has conducted multiple quasi-experimental studies, which are described below.

In the Montessori study that most closely approximates a randomized design, Lillard et al. (2006) investigated self-regulation outcomes for 30 Montessori preschool children and 29 control group children. Although school district policies prohibited the researchers from randomly assigning students, the study relied upon the existing lottery program that assigns students in the relevant school district to schools. Once students were assigned to schools, the researchers assessed children's self-regulation using the Dimensional Change Card Sort (DCCS) task at the end of the school year.

The results indicated significant ($t_{[53]} = 2.11, p < .05$) post-test gains for Montessori students relative to control group students on the DCCS task. However, the study had two key flaws: 1) a small sample, and 2) unreliable randomization process. Firstly, the study tracked only 30 Montessori and 29 control group students, which means that the differences in outcomes could be attributed to the differential teacher quality, school quality, or a host of other factors that affected the small number of students in the sample. Secondly, the study lacks

information about the district and its lottery process, so the quality of the randomization is unclear (Lillard, 2006); thus, it is further unclear whether the Montessori gains resulted from program quality or other unobserved factors.

In another study, Lillard (2012) found significant gains ($d = .53$, $p < .01$) for Montessori children's Heads-Toes-Knees-Shoulders (HTKS) scores relative to control group students. However, this study did involve not any form of random assignment, and the sampled children were all tuition-paying private school students. Over 90% of the sampled parents were white and highly educated, and all parents were fluent in English (Lillard, 2012, p. 386). Thus, issues with sample representativeness as well as issues with the study design (i.e., no random assignment to treatment and control conditions) call the results into question.

Once again, no Montessori evaluation studies exist with randomized designs (Lillard, 2005), and no early childhood studies from outside of the United States context were recovered in the search. The existing studies from the United States have quasi-experimental study designs without statistical controls for potential confounds, which precludes the establishment of causality regarding the effect of Montessori on children's self-regulation.

Other researchers (Camp, Judge, Bye, Fox, & Mattern, 1997; Rathunde, K. Csikszentmihalyi, 2005) have conducted evaluation studies with older Montessori students, but Lillard's work comprises the entire evidence base on early childhood Montessori programs. Given that the preliminary findings from the studies described here (Lillard, 2006, 2012) are promising, more rigorous evidence would be valuable to substantiate the effectiveness of Montessori in improving children's self-regulation.

7.5 Tools of the Mind

Although the curricula above lack conclusive evidence of self-regulation promotion, none of those curricula claim self-regulation development to be their primary focus. To the best of my knowledge, Tools is the only comprehensive curriculum that identifies self-regulation cultivation as its paramount aim. This distinguishing factor of Tools underscores the research rationale of this dissertation.

That is, if Tools claims to be concerned primarily with children's self-regulation growth, then that claim must be rigorously tested to determine whether the program's widespread implementation should continue. The sections below review gaps in the existing Tools literature and then outline how the remainder of this dissertation attempts to fill those gaps.

7.5.1 Gaps in the existing Tools literature

In order for this dissertation's Tools analysis to be useful, it is important to first understand how Tools has been evaluated in previous studies. The existing literature exhibits multiple evaluation studies with heterogeneous results – some studies indicate positive impacts for Tools students (Blair & Raver, 2014; Diamond, Barnett, Thomas, & Munro, 2007), some with null impacts (Hseuh, Lowenstein, Morris, & Mattera, 2014; Lonigan & Phillips, 2012), and some with negative impacts (Clements et al., 2014; Farran & Wilson, 2014). The systematic review in Chapter Nine will review this Tools research base in detail.

Thus, despite relatively similar research designs (i.e., most of the evaluation studies are randomized controlled trials) with substantial methodological overlap, the results across studies exhibit minimal overlap. The mixed findings from existing Tools studies have thus far precluded any authoritative conclusion regarding the curriculum's effectiveness.

Moreover, the Tools evidence base lacks a research synthesis study (e.g., a systematic review and meta-analysis) that aggregates findings across studies. A research synthesis study would estimate both Tools' overall effect as well as the consistency of Tools' effectiveness across different samples and student demographics. Currently, with a mixed Tools evidence base and without a research synthesis study, education officials and policymakers are left to make Tools implementation decisions without a clear body of evidence.

Secondly, whereas existing Tools evaluation studies have investigated Tools' effectiveness at the curricular level, no study has analyzed Tools' specific learning activities. That is, Tools includes 61 discrete learning activities that collectively comprise the curriculum; while researchers have analyzed Tools' effectiveness as a whole program versus other programs, no studies have analyzed how various Tools activities differentially impact child outcomes.

Although the individual activities have never been empirically analyzed, the Tools curricular developers (Leong & Bodrova, 2011) still assert that “implementing specific instructional strategies comprising this curriculum impact the development of self-regulation/executive functions in young children” (p. 12). Thus, the authors claim that implementation of the specific Tools activities does affect children’s self-regulatory capacity, even though this hypothesis has not been directly tested.

Other researchers have also called for an analysis of Tools-specific instructional activities. For example, although two longitudinal Tools evaluation studies (Blair & Raver, 2014; Farran & Wilson, 2014) reached opposite conclusions regarding Tools’ effectiveness, the authors of both have encouraged future researchers to identify the Tools components that do effectively cultivate self-regulation. As a complex intervention consisting of 61 activities, it is crucial to determine the program elements that enhance children’s development, those that potentially undermine development, and those that have no effect at all. No study has yet addressed this question, which constitutes a substantial literature gap.

7.5.2 How this dissertation addresses the literature gaps

Once again, the Tools literature gaps arise from inconsistency in the research base as well as an absence of research regarding specific Tools activities. Based on these gaps, the present dissertation includes two studies: 1) a systematic review and meta-analysis of the existing Tools literature to provide a more informed conclusion regarding Tools’ effectiveness (see the upcoming Section III of this dissertation), and 2) an analysis of the associations between individual Tools activities and children’s self-regulation (see Section IV of this dissertation).

Neither the research synthesis nor the activity analysis has been conducted in any existing Tools research. Thus, it is hoped that both studies represent meaningful contributions to the literature that will inform researchers, policymakers, and practitioners who are determining whether and how to implement Tools.

SECTION III: METHODOLOGY, RESULTS, AND DISCUSSION FOR STUDY ONE

CHAPTER 8: Methodology for study one

This chapter outlines the methodological approach for the research synthesis of study one, which investigates whether Tools of the Mind (Tools) improves children's self-regulation skills vis-à-vis comparison curricula. Embedded within that overarching research aim are the following three sub-questions:

- What is Tools' aggregate effect size observed across the existing literature?
- Is that aggregate effect size significantly heterogeneous? That is, does Tools' effectiveness vary significantly across the existing studies?
- Can heterogeneity in the aggregate effect size be explained by child- and study-level characteristics? For example, do child characteristics (e.g., free school meal status, gender) explain why Tools is effective in some contexts and not others?

In order to address these questions, this study employs research synthesis methods to both narratively and statistically summarize the existing Tools research base. First, this chapter describes research synthesis methods and justifies their application to study one. Next, this chapter addresses criticisms of research synthesis methods. After that, this chapter provides technical details regarding this study's research synthesis plan in advance of presenting the results in Chapter Nine.

8.1 Rationale for this research synthesis project

Research synthesis refers to the practice of identifying, appraising, and combining the results of previous studies on a similar topic (Littell, Corcoran, & Pillai, 2008). Section 7.5 briefly indicated that existing Tools evaluation research has arrived at varied conclusions regarding the program's efficacy. Thus, this research synthesis plan will synthesize the existing Tools evidence base into a more authoritative, coherent conclusion about Tools' effectiveness.

This research synthesis is especially important given Tools' recent proliferation. That is, Tools' expansion across multiple countries (Bodrova & Leong, 2015a) suggests that educators have implemented Tools in hopes that it will improve students' self-regulation and academic outcomes. However, the previous chapter presented Tools evaluation studies with

positive (Barnett et al., 2008; Blair & Raver, 2014; Diamond, Barnett, Thomas, & Munro, 2007), null (Clements & Sarama, 2012; Lonigan & Phillips, 2012) and even negative (Farran & Wilson, 2014) results for students in Tools classrooms. In the face of this heterogeneous evidence base, practitioners and policymakers lack an authoritative conclusion regarding whether and how to implement the curriculum. Research synthesis emerges as an ideal tool to confront that issue (Lipsey & Wilson, 2001).

8.2 Research synthesis: Systematic reviews and meta-analyses

Whereas research synthesis refers to the practice of identifying, appraising, and synthesizing existing evidence, the method actually involves two “distinct but highly compatible” (Littell et al., 2008, p. 1) elements: systematic review and meta-analysis. Although the two are often used interchangeably in the literature (Higgins & Green, 2011; Liberati, Altman, Gøtzsche, & Ioannidis, 2009), the two methods actually involve different techniques. In the present study, both a systematic review and a meta-analysis will be conducted; thus, the characteristics of each are described in the sections below.

8.2.1 Systematic review

Systematic review refers to the comprehensive search for and evaluation of the existing research base on a particular topic. By contrast, typical literature reviews (often called ‘narrative reviews’) involve a narrative summary of the evidence. The narrative review process “is rarely explicit, so readers may not be able to tell how evidence was weighed and whether conclusions are biased” (Littell et al., 2008, p. 11). That is, because narrative review authors can choose to share certain study details while withholding others, it is possible for author biases to distort the presentation of research findings on a topic.

In comparison to narrative reviews, systematic reviews aim to minimize author bias in the search for and presentation of evidence (Higgins & Green, 2011). Specifically, systematic reviews require, *ex ante*, a research protocol, which details a pre-specified search strategy and data extraction strategy. In so doing, the review authors are barred from any post hoc shifts that could optimize the review’s findings (e.g., omitting certain databases from the search so that unfavorable studies would be excluded from the review).

In many cases, authors publish their protocols online so that the research community can hold authors accountable to their pre-designated choices (Littell et al., 2008). This dissertation's review protocol (Baron, Evangelou, Malmberg, & Melendez-Torres, 2016) has been published online with the Campbell Collaboration⁸ so other researchers can ensure that this review aligns with the protocol. In the event that I, or any author, deviate from the review protocol, the author must justify the shift to avoid accusations of bias.

Once again, to minimize bias, review protocols require the pre-specification of a search strategy and data extraction strategy. The rationale for each is described below.

Pre-specification of a search strategy

The pre-specification of a search strategy requires authors to preemptively indicate which online databases, journals, search terms, and other search filters will be used and why. By specifying precise search criteria and procedures, the identification process of relevant studies is wholly transparent to other researchers. This quality enables replication by other researchers because all the steps in study search, retrieval, and analysis are enumerated in the review protocol (Ellis, 2010). The ability to replicate any systematic search bolsters confidence in the research process and findings (Bryman, 2012), which makes systematic reviews useful to policymakers (Lipsey & Wilson, 2001).

In addition to searching electronic databases, systematic reviews can involve other search strategies that authors pre-specify in the protocol. Additional search strategies include searching relevant websites, contacting experts in the field, and searching gray literature (i.e., unpublished articles) databases (Higgins & Green, 2011). These additional strategies are useful because bibliographic databases may omit certain studies; for example, online searches may miss relevant articles that lack target words in the title and abstract (e.g., a paper that omits the words "Tools of the Mind" from the title and abstract but discusses Tools in the text body), whereas an expert in the field may know the study.

⁸ The Campbell Collaboration is the premier clearinghouse for research synthesis papers in the social sciences. The present study's registration (Baron et al., 2016) can be found at: <http://www.campbellcollaboration.org/lib/project/361/>.

Secondly, searching for gray literature guards against publication bias issues (Bryman, 2012; Punch, 2014). That is, previous literature reviews indicate (Chan, Hróbjartsson, Haahr, Gøtzsche, & Altman, 2004; Williamson, Gamble, Altman, & Hutton, 2005) that studies with statistically significant results are more likely to be published than studies with non-significant results. The unpublished studies without statistically significant results often go unseen by the research community, whereas the published articles with significant results receive notice from the research community.

Consequently, published research may present a biased estimate of an intervention's effect because the estimate is based only on a nonrandom sub-sample of the entire population of relevant studies. Searching gray literature databases and contacting experts in the field regarding unpublished studies can countervail against the publication bias issue.

Pre-specification of data extraction

In addition to pre-specifying a search strategy, systematic review protocols also explicitly designate which types of data from the studies will be extracted for analysis. In so doing, systematic reviews aim to avoid the haphazard presentation of outcome data. In a systematic review, the determination of the target outcome measures occurs before the actual search; thus, the author ultimately has to share outcome data on all pre-specified outcomes, which reduces selective reporting practices.

For example, the present review designates self-regulation and academic achievement as the target data types. However, some of the existing Tools studies may lack data on one or both of those constructs. By pre-specifying the outcome variables of interest, I avoid selective reporting bias (e.g., only reporting on the self-regulation data because the academic achievement results go against my expectations). If I did not report all relevant data specified in the coding sheet of my online protocol, then the research community could hold me accountable using the online published protocol.

Summary

Overall, systematic reviews aim to “impose discipline on the review process. Discipline and transparency combine to minimize bias” (Littell et al., 2008, p. 10). The pre-specification of

the search and data extraction strategies diminish authors' ability to modify their presentation of results according to biases. Indeed, previous comparisons of systematic and narrative reviews indicate that the former are more likely to arrive at reliable and valid conclusions regarding the body of evidence on a topic (Bushman & Wells, 2001; Carlton & Strawderman, 1996).

Whereas systematic reviews involve the search for and discussion of relevant studies on a particular topic, meta-analysis serves to synthesize quantitative results from each paper. The meta-analysis then produces an aggregate effect size estimate of the intervention vis-à-vis the control group (e.g., Tools versus comparison curricula). Thus, the next sections describe meta-analysis techniques and then justify their application to the present study.

8.2.2 Meta-analysis

Meta-analysis refers to the statistical aggregation of effect sizes from multiple studies in order to produce a summary effect size on some topic of interest. For example, the present research synthesis examines all previous quantitative Tools studies, each of which has estimated one or more effect sizes regarding the Tools curriculum. In meta-analysis, each of those effect sizes is combined using a weighting algorithm (see section 8.4.6) to estimate Tools' overall effect across the population of studies. The sections below describe 1) the estimation of the summary effect, 2) moderation analyses, 3) and sensitivity analyses.

Estimation of the summary effect

In any field, different studies may yield different estimates of an intervention effect (e.g., some studies find that a math intervention works whereas others find it does not). Indeed, substantial heterogeneity in the Tools literature has been noted (Farran & Wilson, 2014); such heterogeneity both 1) calls the widespread implementation of Tools into question, and 2) underscores the value of a meta-analysis to more accurately estimate Tools' effectiveness.

But how does meta-analysis provide a more accurate intervention effect? Specifically, meta-analysis relies on probability theory to produce more accurate estimates of a treatment effect than do individual studies. The central limit theorem dictates that repeated sampling from a population yields point estimates that are normally distributed around the true population

parameter (Field, 2013). Thus, just as primary researchers can use repeated sampling to better approximate a population mean, meta-analysts use multiple estimates from different primary studies to “obtain a better picture of the distribution of effects and more precise parameter estimates” (Littell et al., 2008, p. 81).

Intuitively, the aggregate effect size will only be as accurate as its constituent effect sizes. Thus, just as publication bias distorts findings in a narrative review, publication bias is also the “most potent source of bias” (p. 111) in meta-analyses. For example, we could imagine a scenario in which all the published Tools literature indicates only significant and positive results, whereas the unpublished studies indicate null results or negative results; in this case, then Tools would artificially appear to be uniformly and substantially effective in a meta-analysis that includes only the published research. Consequently, the inclusion of unpublished literature is equally crucial for the meta-analysis process as it is for the systematic review process (Higgins & Green, 2011).

It is worth noting that unpublished (and published) studies should only be included in the study if they meet rigorous methodological standards. All meta-analyses, including the one in this dissertation, set inclusion and exclusion criteria (see section 8.4.2) to ensure that only high-quality research results contribute to the summary effect. If the studies in the meta-analysis have low internal validity, then the summary effect size would be inaccurate as well. Thus, only studies that meet strict methodological criteria should be included, whereas the remaining studies should be noted in the review but excluded from the meta-analysis itself (Higgins & Green, 2011).

Once a systematic review identifies all relevant studies that meet the inclusion criteria, the effect sizes from the included studies are aggregated using the inverse variance weighting method (Lipsey & Wilson, 2001). The logic of the inverse variance method is that effect sizes with less error should receive more weight (hence the ‘inverse’ terminology) in the aggregate effect size.

Essentially, studies with larger samples produce estimates with smaller standard errors (i.e., less error and more precision). The larger the sample, the smaller the standard error;⁹ the smaller the standard error, the more precise the effect size estimate; and, finally, the more precision (i.e., the smaller the standard error) of the effect size estimate, the more weight it receives in the computation of the meta-analysis' aggregate effect size.

Ultimately, this process yields a composite effect size that includes all relevant quantitative results across all relevant intervention studies. Instead of producing an equally-weighted average of the results across studies, the inverse variance method accounts for effect size precision before computing a composite effect size. This approach maximizes accuracy in the meta-analytic estimation process (Littell et al., 2008).

Moderation analyses

Not only do meta-analyses enable estimation of a summary effect across studies, but they also allow for moderation analyses (i.e., an examination of how the summary effect varies across child- and study-level characteristics). For example, the present review could examine whether 'Tools' overall effect varies by child-level characteristics such as gender.

In addition to child-level characteristics, the present review could examine whether 'Tools' overall effect varies by study-level characteristics such as location in order to assess whether 'Tools' exhibits varying levels of effectiveness in different countries. By contrast, typical narrative reviews "cannot systematically account for moderators" (Littell et al., 2008, p. 11) because narrative reviews do not involve statistical aggregation of quantitative results.

In theory, moderation analyses enable researchers to identify the population for which the intervention would be most (or least) effective. For example, moderation analysis could reveal that 'Tools' works best for female SEN students under the age of five and worst for EAL boys over six. Consequently, even if the composite effect size were to suggest a null

⁹ The equation for standard error of the mean is the standard deviation divided by the square root of the sample size ($se = \frac{sd}{\sqrt{n}}$). Thus, the larger the sample size in the denominator, the smaller the standard error estimate.

overall Tools effect, moderation analyses could enable the identification of sub-populations for whom the intervention might work.

In order to be reliable, however, moderation analysis requires at least ten studies per moderator (e.g., ten or more studies that report on student gender if gender is to be used as a moderator) to sufficiently power the analyses (Lipsey & Wilson, 2001). Moreover, the specific moderators should be pre-specified in the protocol to avoid Type I error inflation; that is, a surfeit of moderator analyses could result in a statistically significant effect for a moderator simply by chance (Tabachnick & Fidell, 2013).

Consequently, the optimal approach is to select a small set of characteristics that theory suggests could moderate the intervention's effect (Higgins & Green, 2011). The moderation strategy employed in this thesis will be described in section 8.4.8.

Sensitivity analyses

The final quality of meta-analysis to be described in this section is sensitivity analysis. As explained in section 8.2.1, research synthesis requires pre-specified decisions that the researcher executes during the analytic process. Of course, had the researcher decided otherwise (e.g., decided to use child age instead of gender as a moderator), then the meta-analytic results may change. Sensitivity analyses enable the examination of how results could vary if decisions had been made differently. Specifically, the central aim of sensitivity analysis is to assess whether the results are “robust (consistent) under different assumptions” (Littell et al., 2008, p. 129).

For example, if gender significantly moderates the relationship between Tools and self-regulation, then does that relationship change if we add SEN to the model? If gender remains significant with SEN added to the moderation model, then we can be more confident that gender is a consistent moderator of Tools' effect on children's self-regulation. Because meta-analyses involve many pre-specified decisions, sensitivity analysis enables researchers to analyze how their decisions affect (or do not affect) the results.

Summary of meta-analysis

How do meta-analyses and systematic reviews fit together? Once again, *meta-analysis* refers exclusively to the statistical aggregation of effect sizes from previous literature, whereas *systematic review* refers to the comprehensive search for and evaluation of the existing research base on a particular topic. Thus, “meta-analysis can (and should) be embedded in a systematic review to minimize bias” (Littell et al., 2008, p. 2-3). This reduction of bias and enhanced robustness makes meta-analyses “especially attractive as a basis for policy, practice guidelines, and the like” (Lipsey & Wilson, 2001, p. 167).

As the first Tools systematic review and meta-analysis, the present study will hopefully provide useful information to educators deciding whether and how to implement Tools. Before proceeding with this method, however, it would first be prudent to consider drawbacks of research synthesis methods. In so doing, the method’s potential obstacles can be preemptively addressed in this dissertation.

8.3 Drawbacks of research synthesis methods

The drawbacks of research synthesis fall into two categories: 1) problems associated with standardizing the search process, and 2) problems associated with statistically aggregating research findings. Each is considered in turn below.

8.3.1 Problems with standardizing the research process

The first drawback of systematic reviews is that they can fail to capture contextual differences across studies because the search and analysis strategies are highly standardized. For example, the pre-specified data extraction form may guide researchers to code studies according to whether they classify as a randomized control trial (RCT). However, two RCTs may exhibit subtle design differences that systematic reviews fail to capture – one study could be a cluster RCT whereas the other could be a stratified RCT, but the coding sheet would designate both studies simply as an ‘RCT.’ In other words, systematic searches and coding schemes can obscure important features of studies that narrative reviews can more thoroughly highlight.

Although this critique may seem intuitively problematic, this weakness of systematic reviews also emerges as its strength. Once again, systematic reviews seek to “impose discipline” (Littell et al., 2008, p. 10) so that reviewers cannot selectively share study-level information. Indeed, narrative reviews do provide more flexibility in the presentation of various studies’ contexts and characteristics, but that flexibility can also open the door to bias. Once again, previous comparative studies of systematic versus narrative reviews indicate that narrative reviews are more likely to selectively share study-level information in a biased way than are systematic reviews (Carlton & Strawderman, 1996; Williamson et al., 2005).

Moreover, systematic review authors can capture more nuanced study-level characteristics simply by creating a more detailed data extraction protocol. That is, review authors can code their included studies to illuminate methodological differences (e.g., coding for both cluster RCTs and stratified RCTs) as well as contextual differences (e.g., whether the study occurred in a traditional state school or an academy). Thus, by creating a more detailed protocol from the outset, systematic reviewers can provide similarly rich accounts of included studies as narrative reviews while still minimizing bias.

8.3.2 Problems with statistical aggregation of effect sizes

The second critique of research synthesis involves the statistical aggregation of effect sizes from different studies. Specifically, because different studies use different samples, designs, and procedures, it can be precarious to combine diverse findings into a single effect size. Wilson and Lipsey (2001) write that “effect sizes and other such summary statistics produced by meta-analysis are not meaningful if they are aggregated over incommensurable study findings” (p. 8). Thus, if the effects across studies are significantly heterogeneous, then the meta-analytic researcher must exercise caution when determining whether and how to summarize results.

Fortunately, recent advances in meta-analytic techniques enable researchers to quantify the heterogeneity in effects using the Q-statistic (Lipsey & Wilson, 2001). The Q-statistic assesses whether the observed effect size variability across studies is significantly more than would be anticipated by sampling error. If the observed effect sizes are more variable than sampling error would predict, then the researcher can conclude that the differences in effect

sizes must be attributable to some other factor besides random sampling variability (Lipsey & Wilson, 2001).

For example, perhaps Tools was effective in one study and not in another because of a study-level characteristic (e.g., an RCT versus a non-RCT) or a child-level characteristic (e.g., one study had many more EAL students than another). Moderation analysis can then be used to identify predictors of the observed heterogeneity in order to determine why the program was effective in some contexts but not others. Thus, whereas heterogeneity in the effect size may have initially been problematic, meta-analytic advances have now transformed heterogeneity into an area of substantive interest (i.e., analyzing which study- and child-level characteristics explain that heterogeneity in findings across studies).

In sum, although research synthesis exhibits various weaknesses, the researcher can mostly mitigate the consequences of those weaknesses through responsible execution. The sections below describe the research synthesis plan to be executed in this dissertation.

8.4 The approach of the present review

Section 8.4 outlines the research synthesis approach for the Tools systematic review and meta-analysis in this dissertation. The sections below align with the protocol guidelines of the Campbell Collaboration, which is the preeminent research synthesis body for the social sciences (Doyle et al., 2010).

The Campbell Collaboration for the social sciences grew out of the Cochrane Collaboration for the medical sciences, which originally developed many of the research synthesis methods (Deeks & Higgins, 2010) in contemporary use. The Cochrane Collaboration also codified many of their research synthesis methods in the freely available Cochrane Handbook for Systematic Reviews of Interventions (Higgins & Green, 2011), to which this review will often refer below as a research synthesis guide.

Thus, once again, the sections below adhere to the protocol outline provided by the Campbell Collaboration. The protocol outline includes specific sub-headings to which reviewers are expected to adhere in order to maximize systematization and minimize bias.

8.4.1 Research synthesis objective

My central objective is to identify, appraise, and synthesize the available evidence regarding Tools in order to evaluate Tools' effectiveness compared with other curricula, including 'business-as-usual' and other programs. In addressing this objective, the meta-analysis will produce an aggregate effect size that characterizes Tools' overall effect across all relevant existing studies, which enables a more authoritative conclusion of Tools' effectiveness than do individual studies.

My ancillary objective is to examine study- and child-level characteristics that explain any observed heterogeneity in effect sizes across trials. If a sufficient number of studies for inclusion are identified (i.e., ten per moderator), then the study-level moderators will include:

- *Publication status:* Does the intervention effect systematically vary between published versus unpublished settings? If so, then this would serve as preliminary evidence of publication bias.
- *Research design:* Does the intervention effect systematically vary between experimental and quasi-experimental designs? If the intervention effect is consistently strong in quasi-experimental designs and null in experimental designs, then this could suggest that the significant quasi-experimental findings are actually attributable to some confounding variable that is not controlled for by the analysis.
- *Implementation fidelity:* Does the intervention effect systematically vary across different levels of implementation fidelity? If Tools has a strong positive effect in high-fidelity settings versus a weak or null effect in low-fidelity settings, then this would underscore the importance of implementing the program with fidelity.

The child-level moderators will include:

- *Gender:* Girls generally exhibit more advanced self-regulation than boys (Barnett et al., 2008; Piotrowski et al., 2013), so does Tools' effect vary across gender?
- *Special education status:* Many children with special educational needs exhibit self-regulatory problems (Harris et al., 2005; Soares et al., 2009), so does Tools differentially affect SEN and mainstream students?

- *Age*: Older children typically have more developed self-regulation than younger children (Piotrowski et al., 2013), so does 'Tools' effect vary for kindergarten versus pre-kindergarten students? This moderator will be included as a binary variable (kindergarten versus pre-kindergarten) instead of a continuous age variable in hopes of identifying sufficient studies to include this as a moderator. With continuous age instead of binary age, many more studies with diverse age ranges would be necessary to power the moderation analysis.
- *Free school meals (FSM) eligibility*: Students from low-income backgrounds typically exhibit more self-regulatory problems than do other students (Raver, 2012; Raver et al., 2013), so does 'Tools' effect change across students with and without free school meal eligibility?

8.4.2 Criteria for including and excluding studies

As explained in section 8.2.2, the meta-analysis' composite effect size is only meaningful if it is comprised of results that are reliable and valid. Thus, research synthesis requires specific inclusion criteria to ensure that the included studies meet a certain standard of methodological rigor. If the inclusion criteria are sufficiently rigorous, then the resultant meta-analytic findings are more reliable because the quantitative results that comprise the composite effect size are individually trustworthy. The inclusion criteria are outlined in the sub-sections below.

Types of study designs

Included studies should have experimental, quasi-experimental, or non-experimental designs that have mechanisms to control for potential confounds in the quantitative effect sizes. Thus, qualitative research designs will be excluded from the present review. Instead, the included studies must qualify as one of the following research designs:

- *Randomized controlled trial*: Random assignment of participants to treatment and control groups by the researcher, using a reliable method of randomization (e.g., computer-generated randomization, coin flips, etc.)
- *Regression discontinuity*: Researchers assign a threshold or cutoff point (e.g., a birthday cutoff for eligibility into an early childhood program) above or below which the

intervention is delivered. Although formal randomization does not occur, comparison of observations lying close to either side of the threshold enables accurate estimation of the treatment effect.

- *Matched control group studies:* Treatment group participants are compared against a matched group of controls who are similar on a set of pre-specified characteristics but do not receive the intervention.
- *Time-series:* Participants are observed before, during, and after the intervention to determine whether it had any effect differentiable from underlying trends over time.
- *Pre- and post-design:* The treatment and control groups, although not randomly assigned, are tested at the beginning and end of the intervention. The pre-test establishes whether significant group differences exist at the study's outset; the post-test reveals whether a significant treatment effect manifests.

Although the review would ideally restrict included studies to randomized trials, randomization of students in education research can be difficult given ethical concerns and school district policies. Thus, this review will also include the quasi-experimental designs described above as long as those studies' designs enable controlling for potential confounds. Moreover, the moderation analyses in the results section (Chapter Nine) will assess whether the results vary between experimental and quasi-experimental designs.

Types of participants

Students of any age, gender, special education status, language learning status, and socio-economic status will be included in this review. This is because the review's aim is to estimate an overall Tools effect for any students who experience the curriculum. Given that Tools is designed as an early childhood curriculum for mainstream classrooms, I expect that included studies will have students of varied backgrounds. The moderation analyses enable assessment of whether the intervention effect varied across those student sub-groups (i.e., FSM, gender, and SEN specifically).

Types of interventions

Any study that analyzes Tools' effect in comparison to one or more 'business-as-usual' or other programs will be included in this review. 'Business as usual' curricula are those that

the school has used before the intervention study began. I will also include studies where Tools was compared with a non-‘business as usual’ program (i.e., where the control group school is also implementing a new educational program or intervention). Sensitivity analyses will be conducted (section 8.4.9) to assess whether the results change when including or excluding non-business-as-usual (i.e., other newly implemented) programs or curricula. Studies that do not pertain to the Tools curriculum will be excluded from the review.

Types of outcome measures

As indicated in Chapter Two, Tools aims to simultaneously hone children’s self-regulatory and academic skills. Thus, the target outcome measures below are based on the stated goals of the Tools curriculum (Bodrova & Leong, 2007), which involve both self-regulation and academic skills. To be eligible for inclusion in this review, studies must include at least one quantitative outcome pertaining to at least one of the four domains below:

1. **Children’s self-regulation as reported by teachers, school administrators, parents, and/or researchers:** These reports derive from observation periods during which a researcher or teacher rates the child’s self-regulatory behaviors. For example, parents, teachers, or researchers can fill out the Behavioral Rating Inventory of Executive Function – Preschool (BRIEF-P) rating form (Gioia, Espy, & Isquith, 2005), which has 63 items to assess children’s inhibitory control, attentional flexibility, working memory, and overall executive control. Any such self-regulation report instruments, whether standardized (e.g., BRIEF-P) or unstandardized (e.g., report forms created by the schools), will be included in the meta-analysis. Sensitivity analyses will be performed (see section 8.4.9) to assess the robustness of findings across standardized and unstandardized rating forms.
2. **Children’s self-regulation skills as indicated by task-based measures:** These scores derive from children’s task performance on an executive function exercise. For example, the Heads-Toes-Knees-Shoulders task (Ponitz et al., 2009) involves touching the correct body part based on the teacher’s instructions, which change after each round. This activity engages multiple aspects of executive function: 1) *working memory* (remembering the teacher’s directions and acting upon them), 2) *cognitive flexibility* (switching among the rules as they change during each round), and

- 3) *inhibitory control* (not touching the body part that you hear, but rather the body part that the teacher has previously specified through a rule).
3. **Children's literacy skills:** Specifically, any literacy scores on preschool achievement tests (whether standardized or unstandardized) will be included. As with the self-regulation rating scales, sensitivity analysis will be performed to check the robustness of findings across standardized and unstandardized test forms.
 4. **Children's numeracy skills:** Specifically, any numeracy scores on preschool achievement tests (whether standardized or unstandardized) will be included. As with the literacy scores, sensitivity analysis will be performed to check the robustness of findings across standardized and unstandardized test forms.

Duration of follow-up

Any follow-up post-test data from the original studies will be included in this review. In line with the Cochrane Handbook's (Higgins & Green, 2011) recommendations, the follow-up data will be classified into three categories: short-term (i.e., data taken between the end of the Tools intervention year to five months following the intervention), medium-term (i.e., data taken between six and 11 months after the end of the Tools intervention), and long-term (i.e., data taken at 12 months or more after the end of the Tools intervention).

Types of settings

I will include studies from any setting where Tools was implemented. Because Tools is a school-based curriculum, the search is anticipated to yield only school-based studies. Nonetheless, no *a priori* setting-based exclusion criteria will be imposed. As long as the study meets the methodological criteria outlined in section 8.4.2, the study setting will not affect the decision to include or exclude the study.

8.4.3 Search strategy

The search strategy designed for this study includes four components recommended by the Cochrane Handbook for Systematic Reviews of Interventions (Higgins & Green, 2011). This four-pronged strategy aims to maximize the comprehensiveness of the search. The four search elements require me to:

- Systematically query the list of databases below. For each database, some variant of “Tools of the Mind” will be used as a search term. For example, in the ERIC database, I will use the following search term: TI(“Tools of the Mind”) OR AB(“Tools of the Mind”). This approach aims to capture every study that mentions Tools at any point in the title (TI) or abstract (AB). The full set of databases to be searched in this review includes:
 - Applied Social Sciences Index and Abstracts (ProQuest)
 - CENTRAL (Cochrane Library)
 - Embase (Ovid)
 - ERIC (ProQuest)
 - LILACS (<http://lilacs.bvsalud.org/en/>)
 - MEDLINE (Ovid)
 - OpenGrey (www.opengrey.eu/)
 - PsycINFO (Ovid)
 - ProQuest Dissertations and Theses (ProQuest)
 - Social Sciences Citation Index (ProQuest)
 - Sociological Abstracts (ProQuest)
- Examine the reference lists of relevant primary studies and reviews to identify additional articles.
- Review the websites of education institutions and technical agencies including:
 - *Tools of the Mind* website: (<http://www.toolsofthemind.org>)
 - *What Works Clearinghouse* at the Institute of Educational Sciences: (<http://ies.ed.gov/ncee/wwc/>)
 - National Institute for Early Education Research: (<http://nieer.org>)
 - Peabody Research Institute: (<http://peabody.vanderbilt.edu/research/pri/>)
 - Society for research on educational effectiveness: <https://www.sree.org/>
- Contact experts in the field to inquire about ongoing studies, gray literature, and potential additional contacts.

8.4.4 Details of eligibility screening

I will independently conduct eligibility screening on all retrieved studies. Specifically, I will screen titles, abstracts, and (where appropriate) full texts in order to determine whether studies are suitable for inclusion in the review. As for coding the study-level and child-level characteristics, I have developed a data extraction form for this review (see Appendix A for the coding form from my online published protocol). I will independently code the studies selected for inclusion into an Excel spreadsheet. In instances of missing or unclear information, I will contact study authors for clarification.

8.4.5 Risk of bias

In addition to coding study characteristics, I will also independently code each study for risk of bias using the Cochrane framework (Higgins & Green, 2011). For example, even though several studies may technically share RCT designs, there may be subtle variability in those research designs that could lead to bias. By systematically evaluating the bias of each study, the overall quality of the research base can be more rigorously assessed. I will rate risk of bias as low-, high-, or unclear-risk across the following six categories as dictated by the Cochrane Handbook (Higgins & Green, 2011):

- *Random sequence generation:* How was random assignment executed? That is, was it done via a computer, a deck of cards, a coin, or some other random generator? If the study claims random assignment but does not explain the assignment mechanism, then this could be a source of bias because participants may not have been truly randomly assigned.
- *Allocation concealment:* Who completed the random assignment? Specifically, did the person who conducted the assignment know which participants were being allocated to which group? If so, then the person might have assigned certain participants to an intervention in a non-random way (e.g., a teacher knew that a child liked dramatic play, so the teacher put that child in the Tools group). By ensuring that the allocation of participants to the experimental group was concealed from the researcher, selection bias can be minimized.

- *Blinding of participants and personnel:* Do the participants know they are receiving the treatment or control? In medical research, the ‘placebo effect’ is avoided by providing each participant with a medication, where one is the true medication and one is a placebo pill. In the present context, it would be impossible to blind teachers and students to the curriculum they are using. Thus, although there is no way to avoid this problem, it remains a source of bias; that is, teachers’ behavior could be affected because they know whether they are using the treatment curriculum or the business-as-usual curriculum.
- *Blinding of outcome assessment:* Do the assessors know the condition assignment of the children they are assessing? If researchers are reporting on children’s self-regulatory behaviors, then it is critical for them not to know whether the students are in a Tools or non-Tools classroom. If researchers know that the child is in a Tools classroom, then the researchers’ evaluation of the child’s self-regulation could, for example, be positively biased by an expectation that the child will be more self-regulated.
- *Incomplete outcome data:* Has there been substantial attrition from the study? If student outcome data are missing from either group, then the results could be biased by whatever factors explain the missingness. If the missing data derives mostly from, for example, FSM students who have moved residences or SEN students who get pulled out of the classroom for individualized instruction, then the results will not represent the true population of students.
- *Selective reporting:* Have all the outcome measures mentioned in the methodology section been reported in the results section? If the study collects data on certain outcome measures but does not report non-significant results, then the reported results could reflect the authors’ biases regarding which outcomes were worthy to report.

In the eventual synthesis, each study’s risk of bias will be reported using a risk of bias table. Any studies with a high risk of bias across multiple categories will be noted in the synthesis, and sensitivity analyses will be performed to determine how results change with and without the biased studies (see section 8.4.9 for the sensitivity analysis plan).

8.4.6 Statistical procedure: Random effects multilevel meta-analysis

As discussed in section 8.2.2, meta-analysis describes the statistical aggregation of effect sizes from various studies into a single composite effect size. In order to create that composite, traditional meta-analysis involves the selection of one effect size from each study to be included in the meta-analytic composite effect size.

If more than one effect size from any given study were included in a meta-analysis, then this would lead to problems with data dependency. That is, one assumption of traditional meta-analysis (as well as all general linear models such as regression and ANOVA) is that the individual observations are independent of one another (Field, 2013). However, this assumption is untenable when a single study has multiple effect sizes. This is because each effect size derives from data involving the same participants; for example, children's teacher-reported self-regulation scores will be related to children's researcher-reported self-regulation scores because those scores are based on the same set of children.

Given the inherent dependency in the data, the assumption of independence of observations is violated. Some meta-analytic studies surmount this problem by selecting one effect size for each study. However, this approach results in a loss of data because the other relevant effect sizes from the included studies are discarded.

Alternatively some meta-analyses overcome this issue by averaging all the effect sizes within studies before computing the composite effect size, but, again, this leads to a loss of data. For example, if I were to average the effect sizes for the assessor-reported and researcher-reported self-regulation skills within each study, then I would have had one effect size instead of two, which reduces statistical power and increases Type II errors (i.e., failing to detect a statistically significant effect that truly exists in the population).

A modern approach to handle the issue of data dependency is multilevel meta-analysis, which does not treat effect sizes from the same study as independent. Rather, multilevel meta-analysis models estimate the shared variation among effect sizes from the same study (see figure 7).

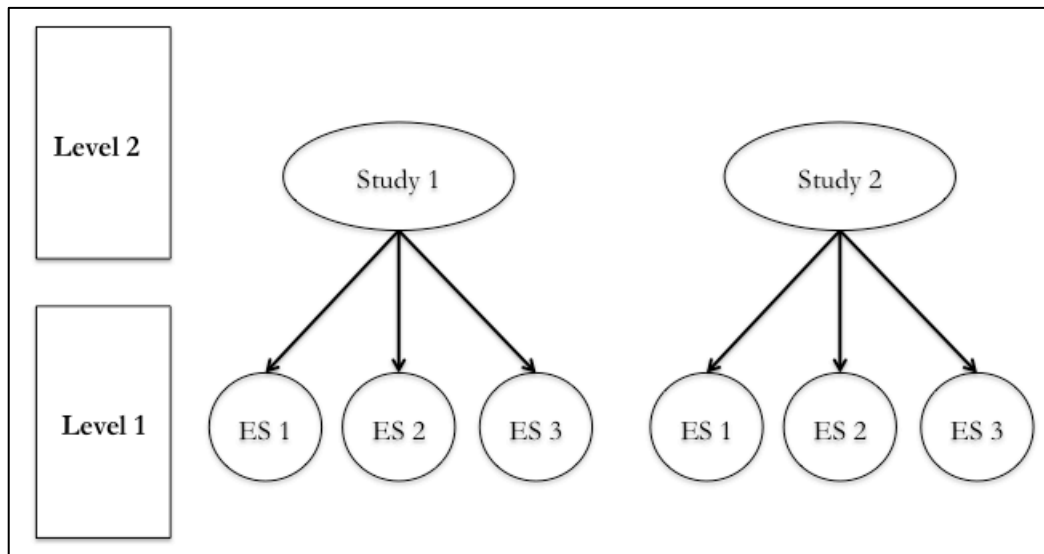


Figure 7: Data structure in multilevel meta-analysis

In figure 7, the individual studies (at level 2) each contain multiple effect sizes (at level 1). The effect sizes in level 1 are not statistically independent from one another because they are nested within the same study based on the same sample of children and same study design. Without accounting for the data dependency among effect sizes from the same study, the standard error of the effect size estimates would be artificially low, which would increase the likelihood of Type I error (i.e., identifying a statistically significant result where no such significant result exists in the real world).

Multilevel meta-analysis, as with traditional single-level meta-analysis, can be conducted as fixed effects or random effects. Both types of meta-analysis produce a composite effect size to summarize the research base. In the fixed effects framework, however, the composite effect size is assumed to apply uniformly to every study included in the meta-analysis. That is, if the observed composite effect size were .3 in favor of Tools, then we would assume that Tools' students' self-regulation skills improved .3 standard deviations relative to the comparison group across all study contexts.

In fixed effects meta-analysis, no study-level heterogeneity is incorporated into the model; instead, fixed effect models account only for the low levels of variability predicted by sampling error. Fixed effects meta-analysis yields more precision in the estimates but, again,

makes more stringent assumptions about the uniform distribution of effect sizes (Lipsey & Wilson, 2001).

Thus, fixed effects approaches are more common in medical literature, where studies generally have lower levels of heterogeneity than in the social sciences. This is because, for example, medical trials where one drug is administered to multiple blinded samples often yield more similar estimates across studies than education trials where one curriculum is implemented in multiple schools.

By contrast, random effects models anticipate some heterogeneity in effect sizes across studies given the differences across studies. Specifically, random effects models anticipate “random variability at both the study-level (studies sampled from a population of studies) and the subject-level (subjects in each study sample from a population of subjects)” (Lipsey & Wilson, 2001, p. 118). Given that Tools is a curriculum that will be implemented among different types of children with different teachers in different schools, it is likely that between-study heterogeneity will emerge.

Thus, for this study, I will conduct a random effects multilevel meta-analysis. In the multilevel approach, data dependency among the effect sizes from the same study is incorporated into the model to adjust the standard errors toward a more accurate estimate. The equation for the random effects multilevel meta-analysis is below:

$$y_{ij} = \mu + w_i + u_{ij} + e_{ij}$$

In the equation above, y_{ij} is an individual effect size (level one in the multilevel structure) from study j (level two in the multilevel structure). Thus, the notation of y_{ij} indicates that the effect sizes are nested within the studies. Next, μ is the estimated grand mean effect size across the entire set of studies. After that, w_i represents the random effect of the heterogeneity between clusters (i.e., estimates the variability in Tools’ effect across studies), whereas u_{ij} represents the random effect of the effect size heterogeneity within a cluster (i.e., estimates the variability across multiple effect sizes from the same study). Finally, e_{ij}

represents the random sampling error that would be expected in any effect size estimation (e_{ij} is the only error term contained in the fixed effects approach).

Because the present meta-analysis will include several studies with several relevant effect sizes in each, random effects multilevel meta-analysis provides the most accurate and least biased summary estimate of 'Tools' effectiveness vis-à-vis comparator curricula. The section below describes how multilevel meta-analysis was executed using statistical software.

Execution of multilevel meta-analysis with statistical software

I will use the metafor package (Viechtbauer, 2010) in the R Studio software to perform random effects multilevel meta-analysis (see Appendix B for sample syntax and output). The metafor R package has been recommended by Wilson (2016), who co-authored the seminal *Practical meta-analysis* manuscript (Lipsey & Wilson, 2001) that outlined many of the research synthesis practices currently used in the social sciences. In addition to multilevel meta-analysis in R Studio, I also used the robust variance estimation (RVE) SPSS (IBM, 2012) macro described in Tanner-Smith & Tipton (2014) to check the robustness of the findings with a slightly different analytic approach (see Appendix B for RVE syntax).

Similar to multilevel meta-analysis, RVE accounts for dependency by adjusting the standard errors for clustering among effect sizes from the same study. Whereas multilevel modeling explicitly models the clustered nature of specific effect sizes with specific studies, RVE simply increases the standard errors to decrease the incidence of Type I errors (Tanner-Smith & Tipton, 2014). Thus, both analytic approaches are appropriate for addressing the issue of effect sizes nested within studies.

Any observed discrepancies between the multilevel and RVE approach will be investigated in the results chapter. It will be possible to compare the multilevel and RVE results because both generate composite effect size estimates based on individual effect size estimates gathered from the included studies. Effect sizes can take several forms; thus, the section below describes the types of effect sizes to be included in this study.

Effect sizes to be used in this study

I will use the standardized mean difference for continuous outcomes and the odds-ratio for binary outcomes. The most common effect size for continuous data is Cohen's *d*, which takes the difference between two groups' mean scores and then standardizes that difference through division by the pooled standard deviation of the two group means. The equation is as follows:

$$d = \frac{Mean1 - Mean2}{SD\ pooled}$$

Instead of Cohen's *d*, this study will use Hedges' *g*, which includes a small sample correction that computes the composite effect size with less bias when sample sizes are small. When sample sizes are large, the Hedges' *g* and Cohen's *d* estimates converge; however, in case the present review identifies studies with low numbers of students in the sample, then the Hedges' *g* estimate will account for the attendant bias. The Hedges' *g* equation, which is a function of the Cohen's *d* estimate described above, is as follows:

$$g = d(1 - \frac{3}{4(n1 + n2) - 9})$$

Whereas Hedges' *g* will be used for continuous outcome data, an odds-ratio will be used for dichotomous outcome data. It is worth noting that binary outcomes are not expected for this study, since it is rare to report self-regulation or academic data with binary measures. Nevertheless, the odd-ratio will be used in case such binary measures do arise.

Odds-ratios capture the difference in odds for some target binary outcome depending on group assignment (e.g., the odds of being rated mathematically proficient versus not proficient for Tools versus non-Tools students). To do so, the odds-ratio equation computes the probability of being a member of one group (e.g., those rated 'proficient') over the probability of being in the other group (i.e., those rated 'not proficient'). The odds-ratio equation is as follows:

$$OR = \frac{P_{Group1}/(1 - P_{Group1})}{P_{Group2}/(1 - P_{Group2})}$$

After calculating effect sizes across all studies, I will convert all included effect sizes into the most common metric (e.g., standardized mean difference). To do so, I will use the Campbell Collaboration's freely available¹⁰ effect size calculator, which can incorporate effect size data on both continuous and nominal scales.

Effect sizes across different comparison conditions

I will compute effect sizes for each experimental condition (e.g., Tools, business-as-usual, other intervention, etc.). For example, if a study compares Tools with another intervention as well as a business-as-usual curriculum, then each of the two comparison conditions will have its own effect size for this review. In so doing, I can incorporate all effect sizes that compare Tools with another program, regardless of whether it is a business-as-usual program or another intervention. The sensitivity analysis will investigate whether the results differ when Tools is compared to a business-as-usual program versus another intervention (section 8.4.9.)

8.4.7 Heterogeneity analysis

I will assess heterogeneity across studies using the Q-statistic. The equation for the Q-statistic is as follows:

$$Q = \sum w_i(T_i - \bar{T})^2$$

In the equation above, w_i denotes the weight of an effect size T_i . The weight is determined by the inverse variance weighting method described in section 8.2.2. An effect size T_i is a single observed effect size from a study, whereas $\bar{T}$ signifies the composite effect size computed across all studies. The resultant Q-statistic indicates the level of heterogeneity observed across studies in the findings (e.g., did different studies arrive at substantially different conclusions regarding Tools' effectiveness?).

¹⁰ The calculator can be found at (http://www.campbellcollaboration.org/resources/effect_size_input.php)

The Q-statistic follows a chi-squared distribution, which enables the assessment of whether statistically significant heterogeneity exists across the effect sizes. If the Q-statistic is significant, then this means that the heterogeneity between the studies' findings exceeds the levels anticipated by sampling error. In that case, the sources of that between-study heterogeneity can be explored using moderation analysis, as described below in section 8.4.8.

8.4.8 Moderation analysis

If a sufficient number of studies (i.e., 10 or more per moderator) are identified, then I will conduct sub-group analyses to determine whether the intervention effect significantly varies across study-level or child-level characteristics. I will execute moderation analyses by comparing the confidence intervals in each sub-group score, which, if not overlapping, signify significant differences across groups.

Once again, the study-level moderators will include:

- *Study design*: Do experimental and quasi-experimental designs exhibit consistently different effect sizes and significance values?
- *Publication status*: Does the intervention effect systematically vary between published versus unpublished settings?
- *Implementation fidelity*: Does the intervention effect systematically vary across different levels of implementation fidelity?

The child-level moderators will include:

- *Age* (percentage of students in pre-kindergarten versus kindergarten)
- *Gender* (percentage of boys versus girls in the study)
- *Special education status* (percentage of SEN students versus non-SEN)
- *Socio-economic background* (percentage of students eligible for free school meals (FSM) versus non-FSM eligibility)

8.4.9 Sensitivity analysis

As described in section 8.2.2, the researcher makes several decisions when outlining the research synthesis protocol. For example, the researcher decides which types of studies to include, how to calculate effect sizes, and so on. Sensitivity analysis enables the researcher to assess the robustness of the findings under different assumptions. This meta-analysis contains the following four sensitivity analyses:

- **Risk of bias:** The meta-analysis will be run both with and without studies that exhibit a high-risk of bias. That is, although a study's research design may dictate its inclusion in this meta-analysis, it may have other risks of bias (see section 8.4.5) that threaten the study's internal validity. Once again, the composite effect size is only as valid as the individual effect sizes that collectively comprise it. Thus, the composite effect size including all studies will be compared against a composite effect size that meta-analyzes only studies without a high-risk of bias.
- **Standardized and unstandardized measures:** In an effort to include as much existing data as possible, this review opted to include both standardized and unstandardized measures of self-regulation and academic skills (see section 8.4.2). Because the reliability and validity of unstandardized test forms are often unclear, I will conduct a sensitivity analysis to estimate a composite effect size with and without the unstandardized test forms.
- **Tools as a stand-alone versus combined intervention:** I expect that some studies will test Tools by itself against a comparison curriculum, whereas other studies will test Tools as a combined intervention (e.g., Tools as a supplement to an existing literacy curriculum). The combined intervention may be implemented substantially differently from the Tools-by-itself program, but those differences may be difficult to discern from the study alone. Thus, I will conduct a sensitivity analysis to estimate a composite effect size both with and without studies that implemented Tools as part of a combined intervention.
- **Tools compared against business-as-usual versus other interventions:** I expect some studies to test Tools against business-as-usual curricula (i.e., the programs that had been in place before the study began), whereas other studies will test Tools

against another newly implemented intervention. Because teachers will likely be more familiar with business-as-usual curricula compared with new intervention curricula, those two curricular types cannot be considered interchangeable. Thus, I will estimate a composite effect size both with and without studies that include comparison programs represented by newly implemented interventions.

Taken together, these four sensitivity analyses assess the robustness of the overall findings in cases where the decision to include or exclude a study could have easily been otherwise (i.e., a study could have been excluded for high-risk of bias or for using an unstandardized assessment form). Because this study aims to include as much available data as possible, it was decided to include all methodologically rigorous Tools studies with quantitative data; however, through the aforementioned sensitivity analyses, it is possible to impose additional rigor on the meta-analytic process to corroborate the robustness of the overall findings.

8.4.10 Publication bias assessment

If a sufficient number of studies (more than 10) are identified, then publication bias will also be visually inspected using a funnel plot. A sample funnel plot is depicted in figure 8. Funnel plots depict the treatment effect estimates as a function of sample size. Given the inverse variance method described in section 8.2.2, large sample studies should produce estimates that are near the true underlying population effect, which is represented by the vertical line in the middle of figure 8.

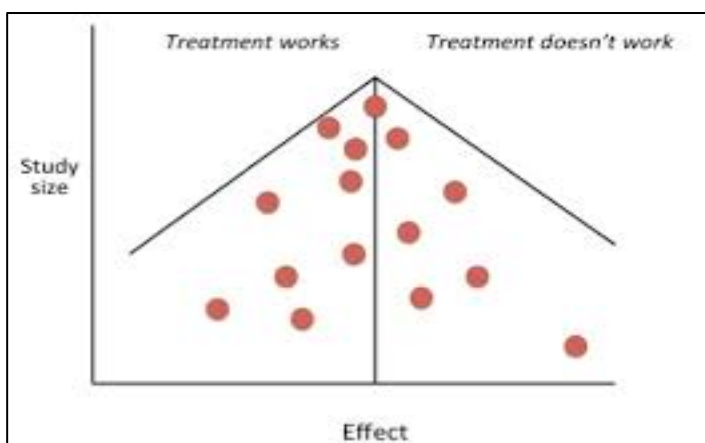


Figure 8: Example funnel plot with no evidence of publication bias

Figure 8 indicates that studies with larger sample studies (i.e., the dots higher up on the y-axis) do cluster tightly around the vertical black line representing the true population effect. Conversely, smaller sample studies (i.e., those lower on the y-axis) are more likely to produce estimates that deviate from the true population effect. That deviation should be a random product of sampling error; consequently, if no publication bias exists, then the plotted treatment estimates should disperse around the true population effect in the shape of the symmetrical funnel depicted in figure 8.

By contrast, if publication bias exists, then the estimates on one side of the funnel would be missing, and the funnel would be non-symmetrical. The funnel plot in figure 9 suggests publication bias because studies with small samples (i.e., those near the bottom of the y-axis) are all on the side indicating that the treatment works. By contrast, truly random sampling error would predict that small sample studies would have a more even distribution of effect size estimates on both sides of the vertical line. Thus, with ten or more studies, funnel plots provide a useful visual mechanism to assess publication bias.

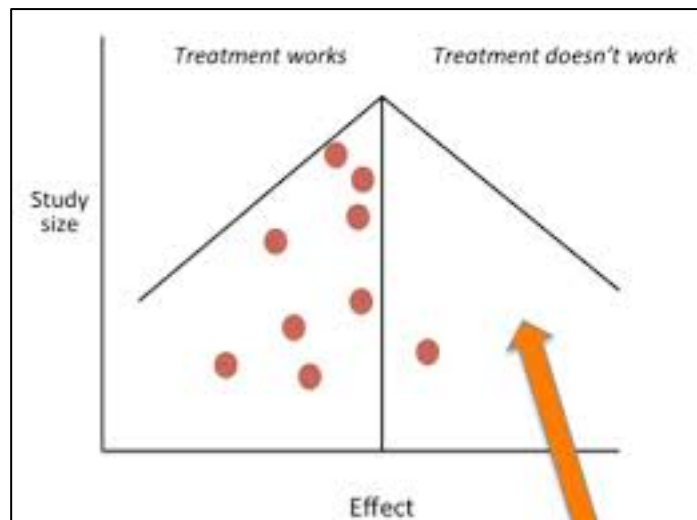


Figure 9: Funnel plot with evidence of publication bias

8.4.11 Ethics for research synthesis

Although the study proposed here does not collect any new data from children, families, or schools, it remains necessary to gain ethical approval (Howe & Moses, 1999). Thus, before

beginning the study, ethical approval was obtained through the Oxford University Departmental Research Ethics Committee. The ethical approval is attached in Appendix C.

In addition to the ethical approval for this study, each of the studies included in this research synthesis received ethical approval from the universities in which the researchers work. The second study of this dissertation also has its own ethics section detailed in Chapter Eleven. Because both studies in this dissertation employ pre-existing anonymous student data, and because the Oxford University Departmental Research Ethics Committee approved both studies, it is hoped that the ethical considerations of this research program have been adequately accounted for.

8.4.12 Summary of research synthesis plan

The present study's research synthesis plan involves both a Tools systematic review and meta-analysis. In order to be included in the quantitative research synthesis, studies must have a research design that enables statistical controls for potential confounds. For example, random assignment ensure that observed effect sizes for Tools versus comparison group students are due to their group assignment instead of unobserved variables.

Once all included studies have been identified through a four-pronged search strategy, the studies will be coded across the categories specified in the Appendix A coding form. In addition to coding each study, the quantitative results from each study will be extracted, and a multi-level meta-analysis will be executed to compute a composite effect size for Tools. Moreover, moderation analyses will be conducted to determine the predictors of heterogeneity in the composite effect size.

Taken together, this plan constitutes the first systematic review and meta-analysis of Tools. It is hoped that the results, to be shared in the upcoming chapter, can inform educators considering Tools implementation.

CHAPTER 9: Results for study one

This chapter presents the results for the Tools systematic review and meta-analysis plan outlined in Chapter Eight. Specifically, this chapter presents the results of the systematic review search process, explains which studies have been included versus excluded and why, describes the characteristics of included studies, and then presents the meta-analytic findings. The results are then discussed in Chapter Ten.

9.1 Results of the systematic search

Before sharing the search results, it is first important to clarify terms. This section refers to both ‘studies’ and ‘records.’ In this review, ‘studies’ are defined as research projects that have analyzed the Tools program; some of the studies were large in scope, which led authors to write multiple ‘records,’ or papers, about them. Thus, it is important to differentiate between a study, which consists of the actual research program, and a record, which is a research paper written about the study.

In the present research synthesis, the search of the 11 electronic databases enumerated in section 8.4.3 yielded 63 total records (see Appendix D for the search terms and results for each database). Ten of those records were duplicates; that is, multiple databases yielded the same record during the search strategy. Thus, I removed duplicate records across databases, which resulted in 53 records.

In addition to the electronic database search, three additional records were identified through other components of the search strategy outlined in section 8.4.3 (i.e., reference list searches, contacting experts, and screening relevant websites). Overall, then, the search process yielded 56 records for title and abstract screening.

After screening the 56 titles and abstracts, 31 records were excluded that did not pertain to the Tools curriculum. I read the remaining 25 full texts and determined 14 records across seven studies to be suitable for inclusion in the present research synthesis (see section 9.2 below for the rationale to exclude the 11 records). That is, the systematic search identified seven separate Tools research programs that fit this study’s inclusion criteria; those seven

research programs, each with its own study ID¹¹ (see table 1 below), were detailed in 14 separate papers.

Table 1: Included studies and the records associated with those studies

Studies included for the systematic review	
Study ID	Records
Diamond (2007)	Diamond et al. (2007)
	Center on the Developing Child (2008)
Barnett (2008)	Barnett et al. (2008)
	Stechuk (2009)
Lonigan (2012)	Lonigan & Phillips (2012)
Blair (2014)	Blair & Raver (2014)
	United States Department of Education (2015)
Clements (2014)	Clements & Sarama (2012)
	Clements & Sarama (2014)
Farran (2014)	Wilson & Farran (2012)
	Farran & Wilson (2014)
Morris (2014)	Mattera et al. (2013)
	Hseuh et al. (2014)
	Morris et al. (2014)

Figure 10 below is a systematic review flowchart (Liberati et al., 2009) that visually depicts the search process from beginning to end.

¹¹ The study ID was chosen by the report from which I gained the most information for the research synthesis. For example, whereas both Barnett et al. (2008) and Stechuk (2009) contained information about the same research program, the Barnett et al. (2008) study contained more information to include in the meta-analysis. Thus, I used Barnett et al. (2008) for the study ID, but the decision easily could have been otherwise without any effect on the results.

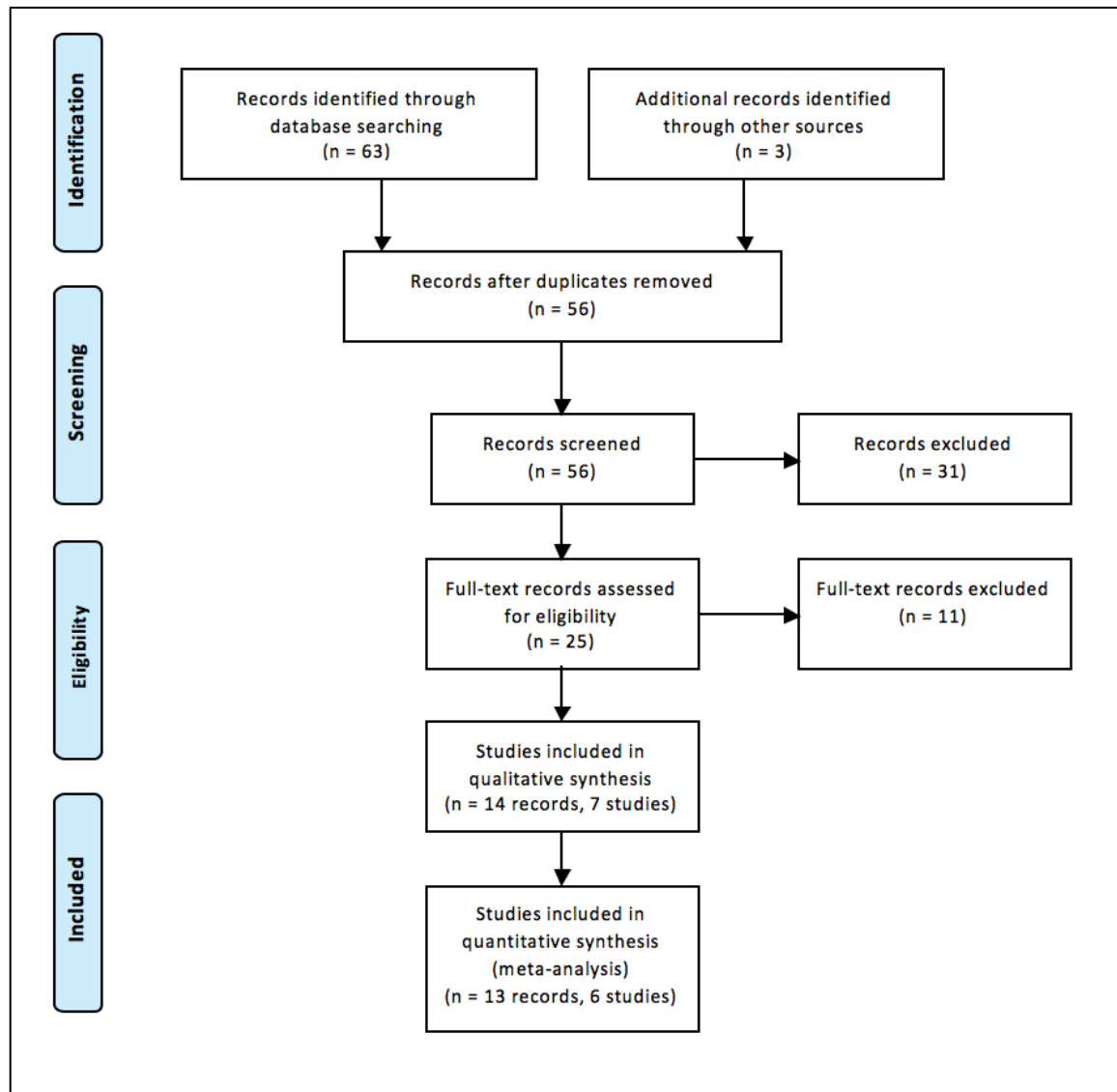


Figure 10: Systematic review flowchart

9.2 Characteristics of excluded studies

As described in section 9.1, 25 of the 56 recovered records were screened in full-text to determine eligibility for the present study. Of those 25 records, 11 were excluded. The 11 excluded studies and the reasons for their exclusion are outlined in table 2 below.

Table 2: Excluded studies and reasons for their exclusion

Study	Reason for exclusion
Bodrova & Leong (2001)	Qualitative study without data on the target outcome measures
Bodrova & Leong (2011)	Theoretical paper with no quantitative data
Copple, C. (2003)	Theoretical paper with no quantitative data
Grigorenko, E. (1998)	Not about the Tools of the Mind curriculum but rather Vygotsky's theoretical ideas
Hammer, E. (2012)	Study has not yet been conducted and thus has not produced any results
Hyson, M. (2006)	A brief qualitative case study of Tools as part of a larger book chapter about child development
Mackay, P. (2013)	Doctoral dissertation with a non-experimental design that did not control for potential statistical confounds
Magalhaes, A. (2013)	Qualitative dissertation without quantitative data on the target outcome measures
Millaway, S. (2015)	Doctoral dissertation with a non-experimental design that did not control for potential statistical confounds
Rodgers, M. (2012)	Qualitative dissertation without quantitative data on the target outcome measures
Shaheen, S. (2014)	Review study with no original quantitative data

9.3 Ongoing studies

The systematic search identified one study (Hammer, Blair, Lopez, Leong, & Bodrova, 2012) that remains ongoing (i.e., the authors have not completed the analysis). Thus, that study's findings could not be incorporated into the systematic review or meta-analysis, but the known features of the study are included in the characteristics of ongoing studies below (table 3). Specifically, the Hammer et al. (2012) study will investigate Tools' impact on students learning English as an additional language.

Table 3: Characteristics of ongoing studies

Study authors	Participants	Intervention	Control	Timeline
Hammer et al. (2012)	420 preschool EAL children, 60 preschool teachers	Tools of the Mind	Not described	Unclear

9.4 Narrative summary of each included study

The salient features of the seven included studies (i.e., not the 14 reports, but rather the seven research programs themselves) are provided below in structured summaries guided by the Participants, Intervention, Control, and Outcomes (PICO) framework used by the Cochrane Collaboration (Higgins & Green, 2011). The PICO framework requires structured information on each study's participants, intervention, control, and outcomes to minimize bias in reporting.

Barnett et al. (2008)

Participants: Barnett et al. (2008) randomly assigned 210 preschool children (54% age four, 46% age three) to either the Tools group (n = 88) or the control group (n = 122).

Intervention: The intervention for Barnett et al. (2008), and in all the sections below, is the Tools program. Since the nature of the Tools program has been extensively discussed in Chapter Two, the sections here do not include additional detail about Tools. Instead, the 'Intervention' portions in the present section (9.4) provide detail regarding the professional development regimen for Tools teachers, the timeline of the study, and information about Tools implementation in the particular study context (e.g., Tools could have been partially implemented in one of the studies, which would be noted in the structured summaries here).

In the Barnett et al. (2008) study, Tools teachers received four days of curriculum training before the start of the school year. During the school year, certified Tools trainers also visited the classrooms once per week. Child-level data were collected during the first year of Tools implementation, so teachers had no training year. This study was conducted nearly ten years ago; thus, Tools had approximately 40 activities (Barnett et al., 2008, p. 301) instead of the 61 activities the program currently includes.

Control: The control classrooms used a curriculum developed by the school district in the three years prior to the study. According to the authors (Barnett et al., 2008), "there was a greater emphasis on teacher-imposed control and less on children regulating each other and themselves" (p. 303). Information regarding the professional development training for control classroom teachers was not provided in the study.

Outcomes: For academic measures, the authors administered the Woodcock Johnson Applied Problems and Letter-Word Identification subtests, the Peabody Picture Vocabulary Test (PPVT – III), the Expressive One Word Picture Vocabulary Test – Revised (EOWPVT – R), the Oral Language Proficiency Test, and the Weschler Preschool Primary Scale of Intelligence (WIPPSI). For self-regulation measures, the authors used the Social Skills Rating Scale (SSRS).

Blair & Raver (2014)

Participants: Blair & Raver (2014) randomly assigned 759 kindergarten children (age statistics not reported) in 79 classrooms in 29 schools to either a Tools group (n = 443) or control group (n = 316).

Intervention: Teachers implemented the Tools curriculum in a two-year professional development cycle. In year one, teachers received five days of training. In year two, teachers received three days of training. Each school also had a Tools coach who provided feedback to teachers once per fortnight in year one and once per month in year two. The intervention was only delivered during children’s kindergarten year of school.

Control: Control group teachers continued business-as-usual practice and professional development training during the two years of the study. According to the study authors, control group classrooms used “commercial literacy and mathematics curricula” that were aligned with state standards (Blair & Raver, 2014, p. 4).

Outcomes: For self-regulation measures, the researchers used the Dimensional Change Card Sort (DCCS), Backward Digit Span, Hearts and Flowers, and Flanker tasks. For academic measures, the authors used the Applied Problems and Letter Word subtests from the Woodcock Johnson III standardized assessment battery as well as the Expressive One Word Picture Vocabulary Test (EOWPVT).

Clements & Sarama (2014)

Participants: Clements & Sarama (2014) randomly assigned 826 children in 84 “four-year-old classrooms” (Clements & Sarama, 2012, p. 2) to one of three conditions: Building Blocks math curriculum, Tools of the Mind plus Building Blocks combined curriculum, or

business-as-usual. Since this meta-analysis pertains to the Tools curriculum, only the Tools students ($n = 288$) and business-as-usual students ($n = 273$) will be referred to hereafter.

Intervention: The study spanned three years. In the first year, teachers began implementing the curriculum to which they had been randomly assigned, but no student data were collected. When the next cohort of pre-kindergarten students arrived in year two of the study, those students were assessed in all measures during the fall and spring of pre-kindergarten. Finally, in year three, all children in all conditions reverted to their business-as-usual curricula (i.e., neither Tools nor Building Blocks were implemented), and follow-up data were collected for all children in all measures. For both years of Tools implementation, teachers received six days of professional development.

Importantly, this study is the only trial in this meta-analysis for which Tools was not implemented on its own in any experimental condition. That is, in this study, Tools was implemented as part of a composite curriculum that included both Tools and the Building Blocks math curriculum. As such, the unique effects of the Tools curriculum vis-à-vis the business-as-usual curriculum cannot be ascertained in this study.

Nonetheless, the inclusion criteria for this meta-analysis outlined in section 8.4.2 dictate that studies involving a combined intervention approach should be included in the research synthesis and then further investigated through sensitivity analysis. Thus, the meta-analysis results in section 9.7 will first include this study in the overall analysis and then compare those results to an analysis in which this study is omitted; any differences between the two results will be reported and discussed.

Control: The control classrooms continued the business-as-usual math curricula used by the three school districts in this study. Specifically, one district used *Everyday mathematics* (McGraw-Hill), one used *Developing math concepts in pre-kindergarten* (from Math Perspectives), and the third had no uniform math curriculum used throughout their schools (Clements et al., 2014, p. 18).

Outcomes: For self-regulation, the researchers used Heads-Toes-Knees-Shoulders (HTKS), Peg Tapping, Forward and Backward Digit Span, Self-Ordered Pointing, and the

Item Selection tasks. For math skills, the researchers used the Tools for Early Assessment of Mathematics (TEAM) and the mathematics portion of the Early Childhood Longitudinal Study (ECLS) cognitive assessment. For literacy, the researchers used Alphabet Knowledge and Name Writing subtests of the Phonological Awareness Literacy Screening (PALS), the Expressive Vocabulary Test (EVT) for vocabulary, and the Refrenew Bus Story to measure oral language and narrative retell.

Diamond et al. (2007)

Participants: Diamond et al. (2007) randomly assigned 147 preschool children (Mean age = 5.1 years) to either a Tools classroom ($n = 7$) or a control classroom ($n = 11$). The Tools children ($n = 62$) and comparison group children ($n = 85$) all attended the same preschool in an urban American city.

Intervention: Intervention group teachers implemented the Tools curriculum with seven days of Tools-specific training during year one and three days during year two. A certified Tools trainer conducted all professional development sessions. Teachers were trained during the first year without any collection of child-level data. In the second year, Tools teachers implemented the curriculum again with a new group of students, whose data were collected for analysis by the researchers. As with Barnett et al. (2008), Tools had approximately 40 activities in this study (Diamond et al., 2007, p. 1387) instead of the current set of 61 activities because the study was conducted ten years ago.

Control: The school district in which the study took place developed a version of the Balanced Literacy (dBL) curriculum. Control group teachers received the same professional development regimen for the dBL curriculum as the Tools teachers did for the Tools program (i.e., seven days during year one and three during year two). A certified school district trainer conducted trainings on the dBL curriculum. According to the authors, “Tools and dBL covered the same academic content, but dBL did not address EF [executive function] development” (Diamond et al., 2007, p. 1387).

Outcomes: Executive function data were collected in the form of both the Dots and Flanker tasks. Children took a pre-test and a post-test for each measure, but the timing of those assessment windows was not reported in either record associated with this study. No

academic achievement data were collected. The actual outcome scores for this study, as well as all the other included studies, will be presented in the multilevel meta-analysis results section (9.7).

Farran and Wilson (2014)

Participants: Farran and Wilson (2014) randomly assigned 877 preschool children (Mean age = 54 months) in 60 classrooms in 59 schools to the Tools condition (n = 646 children) or the control condition (n = 499 children).

Intervention: Teachers implemented the curriculum in a two-year cycle. In the first year, teachers received Tools professional development training (amount is unreported), but no outcome data were collected. In the second year, teachers received more professional development training (amount is unreported), and child outcome data was collected. The Tools intervention was only implemented during children's pre-kindergarten year of school.

Control: Control group teachers continued business-as-usual practice and professional development training during the study. The study took place in five school districts, so "the comparison classrooms used a variety of curricula, with the modal one being Creative Curriculum" (Farran & Wilson, 2014, p. 11).

Outcomes: For self-regulation, the researchers used the researcher-reported Self-Regulation Assessor Rating (SAR) and the teacher-reported Cooper-Farran Behavioral Rating Scale (CFBRS). In addition to those two informant-report measures, the researchers also used the Peg Tapping, Heads-Toes-Knees-Shoulders, Corsi Blocks, Copy Design, and DCCS tasks to measure executive function. For academic skills, the researchers used seven Woodcock Johnson III subtests: Letter Word, Applied Problems, Oral Comprehension, Spelling, Picture Vocabulary, Academic Knowledge, and Quantitative Concepts.

Lonigan & Phillips (2012)

Participants: Lonigan and Phillips (2012) randomly assigned 2,564 children (mean age = 52.7 months, SD = 6.37) in 117 preschool centers to one of four conditions: Tools, Literacy Express Comprehensive Preschool Curriculum (LECPC), a combined curriculum with both Tools and LECPC, and a 'business-as-usual' condition.

Intervention: Teachers in the Tools-only condition implemented the entire Tools program, whereas teachers in the Tools-LECPC combined curriculum only implemented Tools' make-believe play block (see section 2.4.1). Lonigan and Phillips (2012) state that teachers in both Tools conditions received professional development to support "sophisticated and self-regulated play by the children" (p. 3), but the study does not indicate how much training the teachers received.

Each classroom maintained its condition assignment for two years, and data were collected across two sequential cohorts of students for each classroom. That is, each teacher delivered his or her target curriculum over two years with two different groups of students.

Control: Control classrooms continued their 'business-as-usual' practice throughout the two years of the study. Lonigan and Phillips (2012) indicated that 'business-as-usual' classrooms mostly used the HighScope curriculum or Creative Curriculum (see Chapter Seven for more information on these two curricula).

Outcomes: For self-regulation measures, the authors used the Heads-Toes-Knees-Shoulders task as well as the Behavioral Rating Inventory of Executive Function – Preschool (BRIEF-P) to rate children's executive function. For academic measures, the authors used the Bracken Basic Concept Scales – Revised (BBCS – R) as well as the Test of Preschool Early Literacy (TOPEL). The BBCS – R assesses children in six areas: colors, shapes, counting, letters, size, and comparisons. The TOPEL has four subscales: print knowledge, definitional vocabulary, blending sounds, and elisions (Lonigan & Phillips, 2012, p. 4).

Morris et al. (2014)

Participants: 2,670 children in 307 classrooms in 104 preschool centers were randomly assigned to one of four conditions: Tools of the Mind, Incredible Years (IY), Promoting Alternative Thinking Strategies (PATHS), or business-as-usual. All reported comparisons were between an intervention group and business-as-usual; thus, no comparisons of Tools with the Incredible Years or PATHS program were reported. Thus, since this meta-analysis pertains to the Tools curriculum, only the Tools students ($n = 678$) and business-as-usual students ($n = 676$) will be referred to hereafter.

Intervention: Tools training, implementation, and data collection took place in the course of one school year. Nonetheless, the researchers refer to the “comprehensive professional development system for teachers – including four to six training sessions, weekly coaching sessions in the classroom, a ‘real-time’ managing information system (MIS) to support monitoring, and technical assistance” (Morris et al., 2014, p. 2) to support robust implementation across all sites.

Control: The control classrooms continued business-as-usual practice and received no additional professional training above their usual schedule. Of business-as-usual classrooms, 88% used either Creative Curriculum or HighScope.

Outcomes: For self-regulation, the researchers used pencil tapping, the Social Skills Rating Scale (SSRS), the Behavioral Problems Index (BPI), and the Cooper-Farran Behavioral Rating Scale (CFBRS). For academic skills, the researchers used 1) the Woodcock Johnson III Letter Word and Applied Problems subtests, 2) the Academic Rating Scale (ARS) Language and Literacy, Mathematical Knowledge, and General Knowledge subtests, and 3) the Expressive One Word Picture Vocabulary Test (EOWPVT).

9.5 Overview of the characteristics of included studies

The aim of this section is to provide an overall sense of the Tools evidence base by briefly summarizing the methods, geographic settings, data analysis, and Tools implementation features across the seven studies that were narratively summarized in the previous section. Once again, this reporting approach aligns with the guidelines of the Campbell Collaboration and the Cochrane Handbook (Higgins & Green, 2011).

9.5.1 Methods

All seven studies featured cluster randomized controlled trial (RCT) designs, thus meeting the methodological inclusion criteria outlined in section 8.4.2. Quasi-experimental studies were also eligible for inclusion in this review, but no quasi-experimental studies were recovered in the search. For the seven RCT studies, five studies (Blair & Raver, 2014; Clements et al., 2014; Farran & Wilson, 2014; Lonigan & Phillips, 2012; Morris et al., 2014)

used schools as the unit of randomization, whereas the other two studies (Barnett et al., 2008; Diamond et al., 2007) used classrooms as the unit of randomization.

This is because the latter two studies (Barnett et al., 2008; Diamond et al., 2007) each implemented Tools in only one school, so the researchers could only randomly assign classrooms within the individual school. By contrast, the five studies that randomized at the school level had larger samples spread across more schools, which is why they randomized at the school level.

All studies were independent evaluations of the Tools program; that is, the program developers did not oversee any of the studies.

9.5.2 Geographic settings

All studies were conducted in the United States. Although schools in Canada and South America have also implemented Tools (Bodrova & Leong, 2015a), no Tools studies from those regions or any other regions were recovered.

9.5.3 Data analysis

Five of the seven included studies (Barnett et al., 2008; Blair & Raver, 2014; Clements et al., 2014; Farran & Wilson, 2014; Lonigan & Phillips, 2012) used multilevel regression models to analyze child outcomes. By contrast, Diamond et al. (2007) used traditional multiple regression, which did not account for potential data dependencies among students in the same classroom. Finally, Morris et al. (2014) did not report their data analysis strategy.

9.5.4 Intervention implementation

In five of the seven included studies (Barnett et al., 2008; Blair & Raver, 2014; Diamond et al., 2007; Farran & Wilson, 2014; Morris et al., 2014), Tools was implemented as a stand-alone intervention to be compared against comparator curricula.

In another study (Clements et al., 2014), Tools was implemented alongside the Building Blocks math curriculum as part of a combined intervention; thus, it was not clear precisely which elements of the Tools curriculum were incorporated and which were not.

The seventh and final study (Lonigan & Phillips, 2012) included two Tools conditions: one with Tools as a stand-alone program and another where Tools supplemented the Literacy Express Comprehensive Preschool Curriculum (LECPC).

9.6 Risk of bias in included studies

Risk of bias assessment enables evaluation of the existing evidence base's overall quality. All included studies were assessed using the Cochrane Handbook (Higgins & Green, 2011) risk of bias tool, which includes the following six sources of bias:

- *Random sequence generation:* How was random assignment executed? If the study claims random assignment but does not explain the assignment mechanism, then this could be a source of bias.
- *Allocation concealment:* Did the person who conducted the assignment know which participants were being allocated to which group? If so, then the person might have assigned certain participants to an intervention in a non-random way (e.g., a teacher put a child in the Tools group because the child liked pretend play).
- *Blinding of participants and personnel:* Do the participants know they are receiving the treatment or control? In the present context, do teachers and students know whether they are receiving the Tools or comparison condition? If so, then their knowledge that they are in the treatment versus control group could bias their approach toward the study.
- *Blinding of outcome assessment:* Do the assessors know the condition assignment of the children they are assessing? If researchers know that the child is in a Tools classroom, then the researchers' evaluation of the child's self-regulation could, for example, be positively biased by an expectation that the child will be more self-regulated.
- *Incomplete outcome data:* Has there been substantial attrition from the study? If the missing data derives mostly from, for example, FSM students who have moved homes or SEN students who get pulled out of the classroom for individualized instruction, then the results will not represent the true population of students.

- *Selective reporting:* Have all the outcome measures mentioned in the methodology section been reported in the results section? If the study collects data on certain outcome measures but does not report non-significant results, then the reported results could reflect the authors' biases regarding which outcomes were worthy to report.

In line with the Cochrane protocol (Higgins & Green, 2011), studies were characterized as low-risk, high-risk, or unclear risk across each risk of bias dimension. Appendix E includes risk of bias tables for each study that provide textual evidence either from the relevant study report(s) or from my correspondence with the authors to substantiate my risk of bias rating. In addition to Appendix E, the sections below assess the overall risk of bias ratings across the seven studies.

9.6.1 Random sequence generation

Across the seven included studies, five studies were considered low-risk for random sequence generation bias, whereas two studies were considered unclear risk. For the five low-risk studies, four used computer-generated randomization (Barnett et al., 2008; Blair & Raver, 2014; Diamond et al., 2007; Farran & Wilson, 2014), whereas the fifth (Clements et al., 2014) used the circular sampling scheme, which has been shown (Lahiri, 1951) to ensure proper randomization.

The remaining two studies (Lonigan & Phillips, 2012; Morris et al., 2014) did not report their random sequence generation process, which explains their rating of unclear risk. Both studies were reported as randomized controlled trials, so it is likely that both studies either attempted or achieved effective randomization. However, without evidence from the study or authors, the studies' potential for random sequence generation bias remains unclear.

9.6.2 Allocation concealment

Across the seven included studies, four were considered low-risk for allocation concealment bias (Barnett et al., 2008; Blair & Raver, 2014; Diamond et al., 2007; Farran & Wilson, 2014), whereas the remaining three studies (Clements et al., 2014; Lonigan & Phillips, 2012; Morris et al., 2014) were considered unclear risk. Appendix E indicates the textual evidence from

each study that indicates who conducted the randomization. Studies where researchers were not themselves controlling the assignment process received ‘low-risk’ ratings, whereas studies without specific information on the assignment process received ‘unclear risk.’

9.6.3 Blinding of participants and personnel

As with all educational interventions, it is not possible to blind the students and teachers to their curricular assignment. Unlike medical trials, where the treatment drug and control drug can be made to look identical (e.g., both are small white pills), education studies such as this one involve curricular approaches that look very different from one another. Of course, teachers must know what curriculum they are using in order to implement it, which precludes the possibility of true participant blinding.

In instances where blinding of participants and personnel is impossible, the Cochrane Handbook (Higgins & Green, 2011) dictates that the studies should be considered to have an unclear risk of bias. In this instance, I disagree with the Cochrane Handbook because teachers’ knowledge of their curricular assignment could bias the way they teach, interact with students, and so on. Thus, I assert that each study should be rated as having a high-risk of bias on this dimension.

However, despite my disagreement with the Cochrane guidelines, I did commit to using the Cochrane risk of bias protocol, which designates an unclear risk of bias rating for cases such as this one. If I were to deviate from my systematic protocol according to a personal preference (i.e., rating the studies as high-risk instead of unclear risk), then this would represent the precise personal bias issues that systematic reviews aim to overcome. Thus, I chose to adhere to the Cochrane protocol, and all included studies were considered to exhibit an unclear risk of bias on this dimension.

9.6.4 Incomplete outcome data

Across the seven studies, three were considered low-risk for attrition-related issues (Barnett et al., 2008; Blair & Raver, 2014; Farran & Wilson, 2014), two were considered high-risk (Clements et al., 2014; Diamond et al., 2007), and two were considered unclear risk (Lonigan & Phillips, 2012; Morris et al., 2014). The low-risk studies each reported the levels of

missingness, their analyses to address the attrition, and the statistically insignificant differences between the attrited participants and the remaining participants.

The two high-risk studies noted substantial attrition in their study but did not conduct analyses to assess the impacts of the attrition. Once again, Appendix E contains textual evidence from each study to indicate why these studies received a ‘high-risk’ rating. Finally, the remaining two studies provided no information regarding attrition, which explains the ‘unclear risk’ designation given to those studies.

9.6.5 Blinding of outcome assessment

All included studies implemented outcome assessment protocols that aimed to ensure blindness of the assessors to the children’s condition. That is, assessors who filled out observational reports of children’s self-regulation were meant to be blind to the child’s group assignment during the assessment period.

However, as Appendix E indicates, each of the study authors suggested that assessors may have intuited children’s group assignment based on student and classroom characteristics. Thus, although the studies were designed to ensure blinding of outcome assessment, the authors could not guarantee that such blindness had occurred. Thus, each of the included studies received an unclear risk of bias rating for blinding of outcome assessment.

9.6.6 Selective reporting

All seven studies exhibited a low-risk of selective reporting bias. This is because each study reported on all outcomes mentioned in each study’s methodology section.

9.6.7 Other sources of bias

No other source of bias was identified within the included studies. That said, across the entire set of studies, it is possible that the Tools literature base suffers from publication bias. With fewer than 10 studies eligible in the present review, however, a visual inspection of publication bias via funnel plot was not possible (see 8.4.10 for funnel plot description).

Even without the funnel plot, however, there is preliminary evidence of publication bias. Specifically, among the present set of included studies, the three studies to indicate statistically significant positive results for Tools have all been published (Barnett et al., 2008; Blair & Raver, 2014; Diamond et al., 2007), whereas three studies that show null or negative effects (Clements et al., 2014; Farran & Wilson, 2014; Lonigan & Phillips, 2012) have not.

In fact, the only ‘null effects’ study to have been published (Morris et al., 2014) was commissioned by the United States government and was thus published as a government report instead of as a research paper. Thus, no studies that indicate null effects for Tools have been published in peer-reviewed academic journals, even though these studies constitute the majority of the Tools evidence base.

As for other common sources of bias, one strength of the current Tools research base is that the Tools curricular developers (Bodrova and Leong) did not conduct any of the included Tools evaluation studies. A review by Gellis and Reid (2004) found that program developers sometimes have more financial and emotional investment in their programs, which can bias research results. This was not an issue for the included Tools studies because the Tools developers did not conduct any of the included studies.

9.6.8 Risk of bias summary

The risk of bias results are visually summarized in figure 11 below. Figure 11 indicates an unclear risk of bias across all studies for the blinding of personnel, which was not possible, as well as for the blinding of outcome assessors, which was difficult to ensure across studies. Moreover, 29% of the studies (two of the seven) exhibited a high risk of attrition bias because of incomplete outcome data.

Beyond that, none of the studies exhibited a high risk of bias across any of the other Cochrane Handbook’s (Higgins & Green, 2011) risk of bias dimensions. The sensitivity analysis section (9.10) will investigate how the meta-analytic results (see section 9.7 below) change based on the inclusion versus exclusion of studies with a high risk of bias.

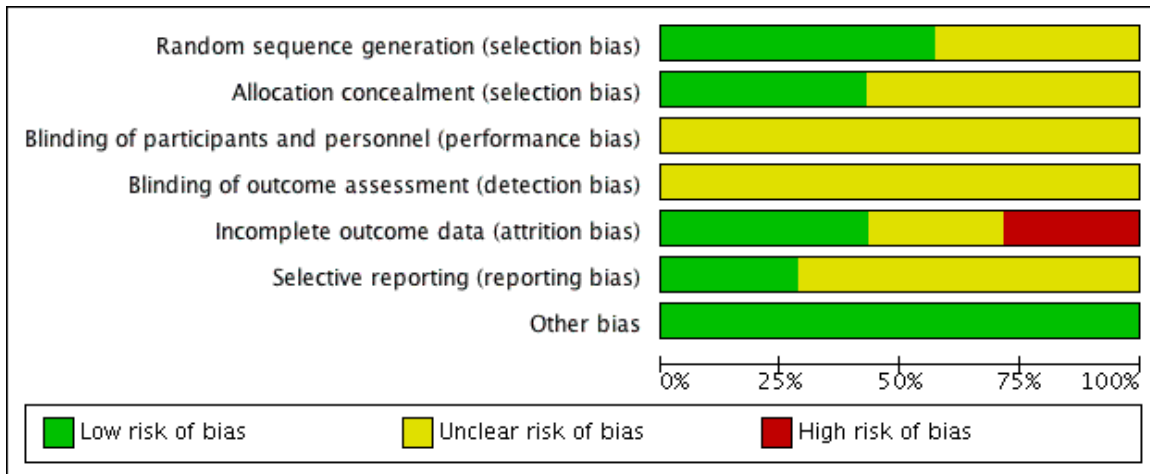


Figure 11: Risk of bias graph presented as percentages representing author's risk of bias judgments

9.7 Multilevel meta-analytic findings on the intervention's effect

The previous sections of this chapter included information relevant for the systematic review portion of the research synthesis study. From here onward, the remaining sections of this chapter detail the results of the meta-analysis, which only concerns the statistical aggregation of effect sizes across studies.

Whereas the systematic review results in the previous sections included information across seven studies, one of the studies (Lonigan & Phillips, 2012) did not report the necessary outcome data to include in this meta-analysis. The authors were contacted numerous times by phone and email, but the authors opted not to share the requisite results, which precluded their study's inclusion in the present review.

Thus, this section presents the meta-analytic results across the six included studies for the quantitative synthesis. The results are divided among the four types of outcome measures described in section 8.4.2:

1. Children's self-regulation as reported by teachers, school administrators, parents, and/or researchers (e.g., BRIEF-P, Social Skills Rating Scale)
2. Children's self-regulation skills as indicated by task-based measures (e.g., Heads-Toes-Knees-Shoulders, Dimensional Change Card Sort)

3. Children’s literacy skills as measured by standardized or unstandardized instruments
4. Children’s math skills as measured by standardized or unstandardized instruments

As described in section 8.4.6, each study yielded multiple effect sizes on at least one of the relevant outcome measures. Because effect sizes from the same study are based on the same sample of children and the same study characteristics, those effect sizes cannot be considered statistically independent from one another. Thus, multilevel meta-analysis was used to account for shared variation among effect sizes from the same study (section 8.4.6).

After incorporating the relevant effect sizes from each study into each outcome, the final results were null across three of the four outcome measures (see table 4). That is, neither the Tools group nor the comparison group performed statistically significantly better than the other for either self-regulation outcome or for literacy. In math, the composite effect size was positive and statistically significant ($g = .051$, $p < .01$) in favor of the Tools condition.

Table 4: Multilevel meta-analysis results across the four outcome measures

Outcome	n(k)	Effect size	SE	p-value	95% C.I.
Reported SR	12(3)	.172	.161	.283	(-.143, .487)
Task-based SR	36(5)	.121	.12	.313	(-.069, .069)
Literacy	35(5)	.021	.029	.486	(-.037, .077)
Math	15(5)	.051	.019	.009	(.013, .091)

(Note: ‘n’ signifies the number of effect sizes; k signifies the number of studies from which those effect sizes were drawn; ‘effect size’ signifies the composite effect size across all studies)

The four sections below further detail the results across the four outcomes. Each section includes a forest plot, which visually depicts the relevant effect sizes gathered from each study as well as the aggregate effect size computed across all studies.

9.7.1 Outcome #1: Children’s self-regulation as reported by teachers, school administrators, parents, and/or researchers

Both teachers and researchers completed ratings of children’s self-regulation skills during the Tools implementation year. For example, the self-regulation assessor rating (SAR; Smith-Donald et al., 2007) requires researchers to observe a child’s classroom behavior and then fill out a five-item questionnaire regarding the child’s self-regulation skills.

These assessor-based ratings came from three of the six studies (Barnett et al., 2008; Farran & Wilson, 2014; Morris et al., 2014), contributing a total of 12 effect sizes for this analysis. In both Barnett et al. (2008) and Morris et al. (2014), the data were collected immediately following the end of the Tools implementation year. In Farran & Wilson (2014), the data were collected at the end of the Tools implementation year (i.e., pre-kindergarten) as well as during the spring of the following two school years.

The forest plot in this section and the three sections below it depict all effect sizes for each outcome as well as their respective 95% confidence intervals. The intervals that cross the vertical line in the forest plot's center indicate no significant effect, whereas error bars that do not overlap with the vertical line indicate a significant effect. Next, the diamond shape and its 95% confidence interval at the bottom of the forest plot represent the composite effect size.

Finally, the percentage values in the columns to the right of the error bars reflect the weight of each effect size within the composite effect size. As described by the inverse variance method section (8.2.2), more precise effect sizes (i.e., those with larger samples and smaller variance values) receive more weight in the composite effect size.

Figure 12 shows that the Social Skills Rating Scale scores for Tools students in Barnett et al. (2008) were positive and statistically significant ($g = .55, p < .05$) in favor of Tools. Beyond that, however, none of the other included effect sizes were significant, and the composite effect size was also relatively small and not significant ($g = .17, p > .05$) in favor of Tools.

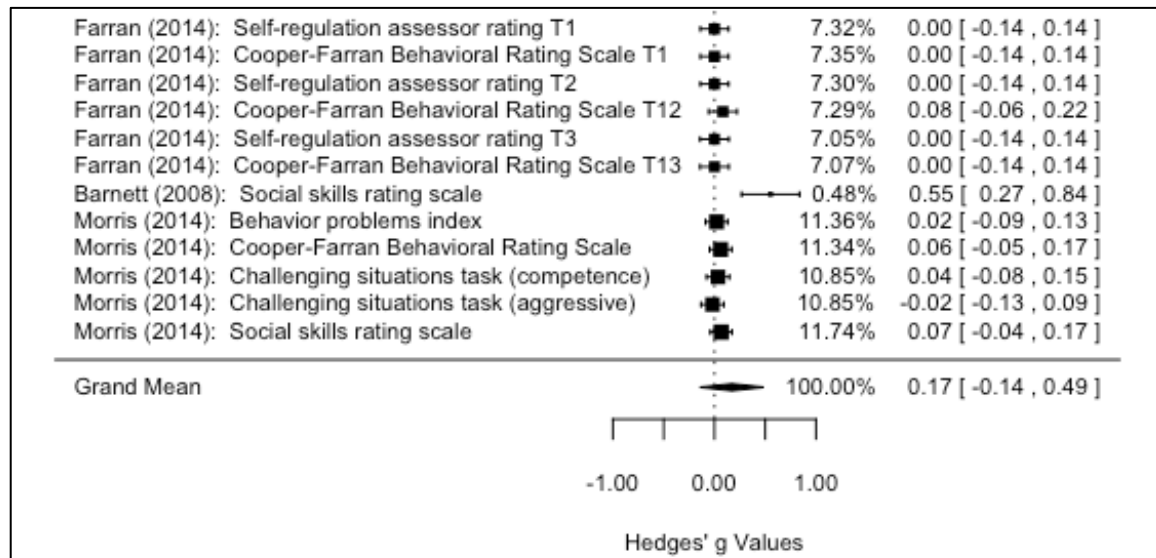


Figure 12: Forest plot for assessor-reported self-regulation effect sizes across studies

9.7.2 Outcome #2: Children's self-regulation skills as indicated by task-based measures

The second outcome was task-based self-regulation measures, where children complete a cognitive control task based on guidance from an adult. For example, the Flanker Task (Diamond et al., 2007) prompts children to press a button on their right when they see a circle and a button on their left when they see a triangle. The target shape is then inscribed within another circle or triangle, and the children receive new rules that dictate when to press one button versus the other (e.g., when the two shapes are congruent, then the children press the left-hand button and vice versa).¹²

These task-based self-regulation data came from five of the six included studies (Blair & Raver, 2014; Diamond, 2007; Clements et al., 2014; Farran & Wilson, 2014; Morris et al., 2014), contributing a total of 36 effect sizes for this analysis. In Blair & Raver (2014),

¹² Tasks such as these tax all three elements of executive function: 1) *working memory* to remember and apply the relevant button-pressing rule, 2) *cognitive flexibility* to switch rule sets across trials (e.g., sometimes a triangle means pressing on the right, and sometimes it means pressing on the left), and 3) *inhibitory control* to suppress the impulse to press a button that would have been correct on a previous trial but is now incorrect based on a new rule.

Diamond (2007), and Morris et al. (2014), the authors collected all task-based data at the end of the Tools implementation year. By contrast, Clements et al. (2014) collected the data at the end of the Tools implementation year as well as one year later for a delayed post-test. Finally, Farran & Wilson (2014) collected the data at the end of the Tools implementation year as well as during the spring of the following two school years.

Figure 13 shows that all effect sizes from the Diamond et al. (2007) study were positive and significant in favor of Tools. Additionally, the Copy Design task from Farran & Wilson (2014) was significant at the first time point in favor of Tools, whereas the DCCS and backward digit span were significant at the second time point in favor of the comparison group children. Beyond those, however, none of the other included effect sizes were significant, and the composite effect size was also relatively small and not statistically significant ($g = .12$, $p > .05$) in favor of Tools.

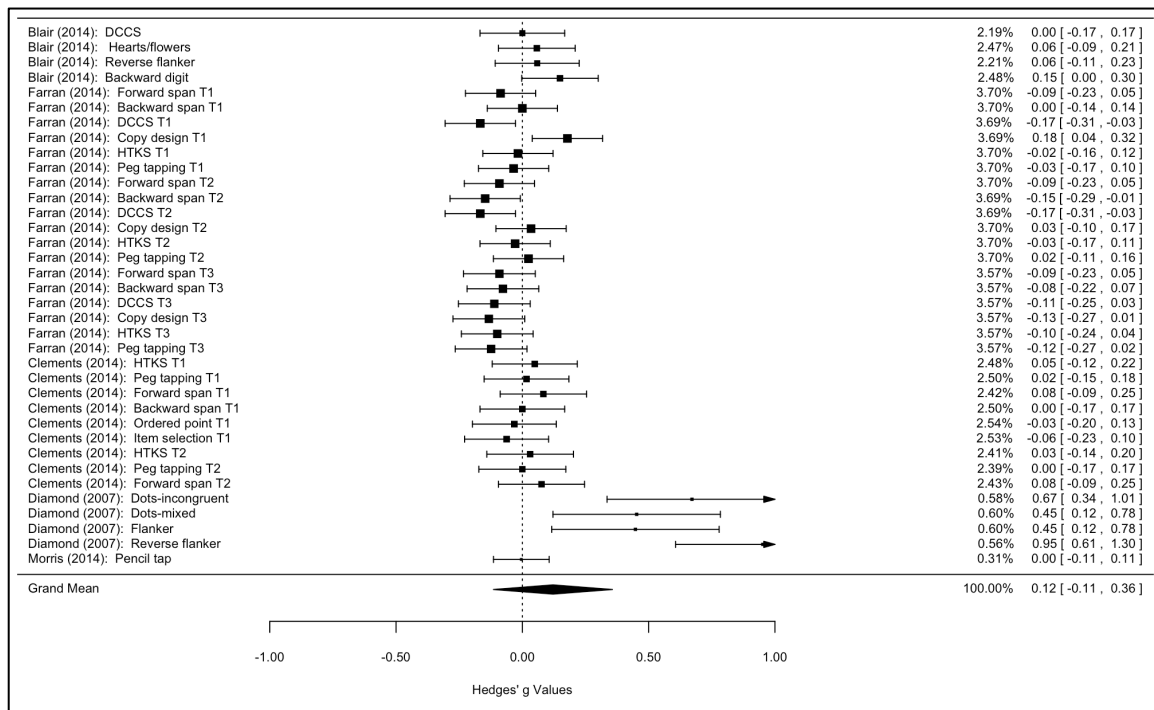


Figure 13: Forest plot for task-based self-regulation effect sizes across studies

9.7.3 Outcome #3: Children's literacy skills

Across the five studies that reported literacy skills (Barnett et al., 2008; Blair & Raver, 2014; Clements et al., 2014; Farran & Wilson, 2014; Morris et al., 2014), children completed a diverse variety of literacy assessments (see section 9.4 for the specific assessments used in each study). In total, these five studies contributed 35 literacy effect sizes.

In Barnett et al. (2008) and Morris et al. (2014), the authors collected the literacy data only at the end of the Tools implementation year. No follow-up literacy data were collected. By contrast, Clements et al. (2014) and Blair & Raver (2014) collected literacy data once during the school year following Tools implementation. Finally, Farran & Wilson (2014) collected the data at the end of the Tools implementation year as well as during the spring of the following two school years.

Figure 14 shows that both vocabulary effect sizes from Blair & Raver (2014) were significantly positive ($g = .16, p < .05$) in favor of Tools. Additionally, the Refrenew bus story complexity subscale from Clements et al. (2014) was significant ($g = -.08, p < .05$) in favor of the comparison group. Beyond those, however, none of the other included effect sizes were significant, and the composite effect size was near zero and not significant ($g = .02, p > .05$) in favor of Tools.

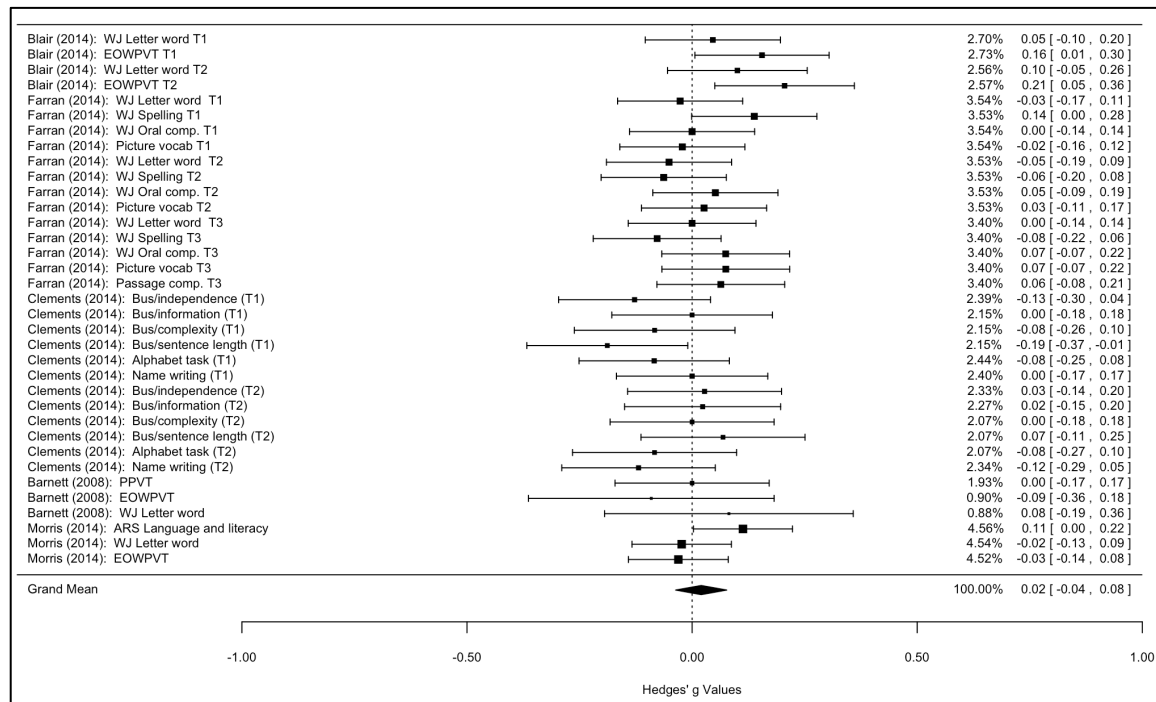


Figure 14: Forest plot for literacy effect sizes across studies

9.7.4 Outcome #4: Children's math skills

Five of the six included studies (Clements et al., 2014; Farran & Wilson, 2014; Morris et al., 2014) collected data on children's math skills, contributing a total of fifteen effect sizes. Morris et al. (2014) collected math data only at the end of the Tools implementation year. Clements et al. (2014) collected math data at the end of the Tools year as well as during the spring of the following school year. And, as with the other data types, Farran & Wilson (2014) collected the data at the end of the Tools implementation year as well as during the spring of the following two school years.

Figure 15 shows that the Woodcock Johnson Applied Problems math subtest results from Blair & Raver (2014) were positive and significant ($g = .11$, $p < .05$) in favor of Tools. Although none of the other included effect sizes were significant, nearly all of them favored the Tools condition. Thus, when combined through meta-analysis, the composite effect size was small but statistically significant ($g = .05$, $p < .01$) in favor of Tools.

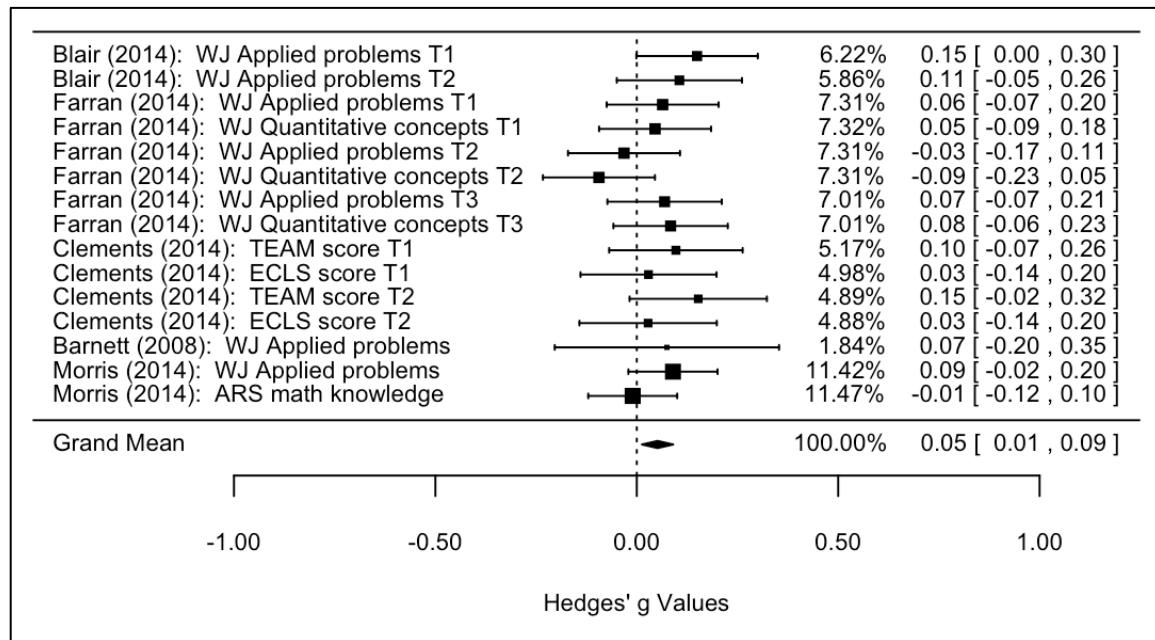


Figure 15: Forest plot for the math effect sizes across studies

9.7.5 Summary of meta-analytic findings across the four outcome measures

The results sections above indicate a statistically significant and positive effect size for math that favors Tools students. For the literacy and two self-regulation composite effect sizes, no statistically or marginally significant results were observed, but all the composite effect sizes favored the Tools condition. These results are discussed further in the upcoming discussion chapter.

9.8 Heterogeneity in the composite effect sizes

In addition to the composite effect size estimates, the heterogeneity of the results was assessed through the Q-statistic. Once again, whereas different effect sizes in different studies would be expected to vary slightly because of sampling error, Q-statistics quantify whether the observed heterogeneity across effect sizes exceeds what would be expected through sampling error.

The results from table 5 indicate significant heterogeneity for the task-based self-regulation composite effect size ($Q = 104.14_{[35]}, p < .001$). Once again, the high heterogeneity value

suggests that the included studies arrived at substantially different conclusions regarding Tools’ effect on children’s task-based self-regulation skills.

By contrast, the remaining three outcome measures exhibited statistically insignificant heterogeneity with small Q-statistic values. Specifically, assessor-reported self-regulation, literacy, and math exhibited mild to moderate shared variance across studies, none of which were significant at $p < .05$. Those null outcomes indicate that the different studies arrived at roughly comparable conclusions for those three outcomes.

Table 5: Heterogeneity analysis for the composite effect sizes

Outcome	Q	p-value
Reported SR	15.85 (11)	.147
Task-based SR	104.14 (35)	.001
Literacy	39.15 (34)	.249
Math	11.44 (14)	.651

The significant heterogeneity for the task-based self-regulation outcome thus merits further moderation analyses, which are discussed below in section 9.9.

9.9 Moderation analysis

Once again, moderation analysis enables the examination of how the summary effect varies across child- and study-level characteristics (e.g., did the high proportion of FSM students in one study explain why that study’s effect size was lower than in other studies?).

Unfortunately, too few studies were recovered to conduct moderation analysis. As section 8.4.8 described, at least ten studies per moderator are necessary (e.g., ten or more studies that report on student gender if gender is to be used as a moderator) to sufficiently power the analysis (Lipsey & Wilson, 2001). The present meta-analysis includes only six studies overall, which precluded moderation analyses.

Nonetheless, as more Tools studies are published in the future, I will be responsible for updating this meta-analysis on the Campbell Collaboration database every three years as indicated by my online protocol (Baron et al., 2016). Once a sufficient number of Tools studies report on certain child-level or study-level characteristics (e.g., almost all studies

report on student gender as well as study design, among other characteristics), I will include moderation analyses in subsequent updates to this meta-analysis.

9.10 Sensitivity analysis

As outlined in section 8.4.9, this study involved four sensitivity analyses. Once again, sensitivity analysis enables the researcher to assess the robustness of the findings when different studies are included or excluded based on the criteria outlined in the protocol. For example, the sensitivity analyses presented here assess how the overall composite effect sizes reported in section 9.7 vary when the following study types are removed from the analysis:

1. Studies with a high risk of bias
2. Studies with unstandardized test forms
3. Studies that test Tools as a combined intervention (i.e., Tools as a supplement to another intervention)
4. Studies that test Tools against a new intervention (i.e., a non-business-as-usual)

Each of these sensitivity analyses is sequentially described in the four sections below.

9.10.1 Studies with a high risk of bias

Although the included studies exhibited low or unclear risk of bias on five of the six Cochrane risk of bias dimensions, two studies had a high-risk of bias on the incomplete outcome data (i.e., study attrition) indicator. Thus, those two studies (Clements et al., 2014; Diamond et al., 2007) were removed from the analysis, and the composite effect size was recalculated using the remaining four studies. In so doing, I could assess whether Tools' impact significantly differed among the studies that met high methodological standards.

After removing the two studies from the analysis, the results remained consistent with the original analysis across the four outcomes (see table 6). Neither of the two excluded studies collected assessor-reported self-regulation data, so results for that outcome were identical. Next, both task-based self-regulation ($g = -.001, p = .879$) and literacy ($g = .041, p = .149$) remained practically and statistically insignificant. Finally, the composite math effect size remained statistically significant ($g = .047, p < .05$) in favor of Tools.

Table 6: Sensitivity analysis to remove studies with a high-risk of bias

Outcome	n(k)	Effect size	SE	p-value	95% C.I.
Reported SR	12(3)	.172	.161	.283	(-.143, .487)
Task-based SR	23(3)	-.001	.044	.879	(-.093, .079)
Literacy	23(4)	.041	.028	.149	(-.014, .095)
Math	11(4)	.047	.024	.047	(.001, .093)

9.10.2 Studies with unstandardized test forms

None of the included studies reported using unstandardized assessment forms for any of the four outcome measures. In each study, the authors provided a reference for the assessment form and provided details of its reliability and validity where relevant. Given that all studies used standardized test forms, I did not conduct a sensitivity analysis on this dimension.

9.10.3 Studies that test Tools as a combined intervention

One included study (Clements et al., 2014) tested Tools as a supplement to another intervention: the Building Blocks math curriculum. Given that the study lacked a Tools-only condition, the unique effect of Tools on children's outcomes could not be assessed. I opted to include studies with combined interventions such as Clements et al. (2014) because I wanted to analyze all the existing Tools data that I could find. Nonetheless, this sensitivity analysis only includes studies that compared Tools on its own against other curricula. In so doing, I was able to investigate whether the overall Tools effect sizes change when Tools is compared as a stand-alone intervention against comparison curricula.

After removing Clements et al. (2014) from the analysis, the results remained consistent with the original analysis across the four outcome measures (see table 7). Clements et al. (2014) did not collect assessor-reported self-regulation data, so those results did not change. As for the other three outcome measures, task-based self-regulation ($g = .15$, $p > .05$) and literacy ($g = .041$, $p = .149$) remained insignificant in favor of Tools, whereas math remained statistically significant ($g = .047$, $p < .05$) in favor of Tools.

Table 7: Sensitivity analysis for Tools as a combined intervention

Outcome	n(k)	Effect size	SE	p-value	95% C.I.
Reported SR	12(3)	.172	.161	.283	(-.143, .487)
Task-based SR	27(4)	.150	.154	.331	(-.152, .453)
Literacy	23(4)	.041	.028	.149	(-.014, .095)
Math	11(4)	.047	.024	.047	(.001, .093)

9.10.4 Studies that test Tools against a new intervention

The only included study to test Tools against another new intervention was the Clements et al. (2014) study described in the previous section. Thus, this sensitivity analysis involved the same process as the previous analysis; that is, I removed the Clements et al. (2014) data and re-ran the composite effect size analysis. Thus, the results of this sensitivity analysis are the same as those reported in the previous section. Specifically, literacy and both self-regulation composite effect sizes remained statistically insignificant, whereas the math composite effect size remained significant in favor of Tools.

9.10.5 Sensitivity analysis summary

Once again, I was unable to conduct the unstandardized assessment form sensitivity analyses because all studies employed standardized test forms across outcome measures. Beyond that, however, the other sensitivity analyses corroborated the findings observed in the original analysis with all available data across the four outcome measures. Thus, it appears that my decisions to include studies with a high-risk of bias, studies that implemented Tools as a combined intervention, and studies that compared Tools with other newly implemented interventions did not significantly alter the results.

9.11 Robustness check with robust variance estimation

As indicated in the previous chapter (section 8.4.6), I assessed the robustness of the multilevel meta-analytic findings using the robust variance estimation (RVE) method. As with multilevel modeling, RVE addresses the issue of data clustering (i.e., when students are nested within the same classroom, or when effect sizes are nested within the same study).

The robustness check results mirror those observed in the multilevel analysis (see table 8). Specifically, the Tools composite effect size was statistically significant for math and null for the remaining three outcome measures, and the effect size magnitudes are highly similar to those observed for the multilevel analysis (see table 4). Thus, this analysis indicates that the findings in the original multilevel meta-analysis are robust to multiple analytic methods.

Table 8: Robustness check with robust variance estimation

Outcome	n(k)	Effect size	SE	p-value	95% C.I.
Reported SR	12(3)	.158	.15	.404	(-.488, .804)
Task-based SR	36(5)	.077	.085	.416	(-.159, .313)
Literacy	35(5)	.027	.027	.369	(-.047, .102)
Math	15(5)	.061	.019	.036	(.001, .115)

9.12 Chapter summary

This chapter presented the results for the Tools systematic review and meta-analysis, both of which are the first of their kind in the literature. This research synthesis recovered a small number of Tools randomized controlled trial evaluation studies, which suggests that more research would be valuable to accurately assess Tools' efficacy.

Across the existing Tools studies, the meta-analytic results indicated significant benefits for Tools students' math skills but not literacy or self-regulation. Thus, although Tools targets children's self-regulation skills, the meta-analytic results did not indicate measurable self-regulatory gains for Tools students.

That said, it is important to note that Tools did not *underperform* relative to the comparison programs on the self-regulation measures; rather, the multilevel meta-analysis and robust variance estimation analyses did not provide conclusive evidence that Tools *outperformed* the comparison programs on the self-regulation outcome measures. In fact, all four composite effect sizes favored Tools (i.e., the beta coefficient was positive instead of negative), even though only one of the effect sizes reached statistical significance. The implications of these findings will now be discussed in the upcoming discussion chapter for study one.

CHAPTER 10: Discussion for study one

This chapter discusses the research synthesis results presented in Chapter Nine. This chapter first summarizes the main results of the systematic review and the meta-analysis. Next, this chapter describes the strengths and limitations of the present study. Finally, possibilities for future research are discussed before transitioning to study two, which analyzes the specific learning activities that collectively comprise Tools.

Traditionally, a discussion chapter would also contain implications for policy and practice. In the present review, however, that section will surface in this dissertation's final chapter, which synthesizes findings from study one (i.e., this research synthesis study) and study two (i.e., the analysis of specific Tools activities). Once both studies have been presented, their collective implications for policy and practice are described in this dissertation's fourteenth and final chapter.

10.1 Summary of the main results

This summary section is divided into three parts: 1) A discussion of the systematic review results, 2) an analysis of potential publication bias in the existing Tools literature, and 3) a summary of the meta-analytic results.

10.1.1 The systematic review results

The present systematic review recovered 14 records across seven studies. Each of those seven studies classifies as a randomized controlled trial (RCT), which is considered to be the 'gold standard' in social science research (Bryman, 2012). No quasi-experimental Tools studies were recovered in the search. Despite the mostly high quality of the included studies, the number of included studies was lower than the required ten for further analyses such as formal publication bias assessment and moderation analysis.

Nonetheless, the small number of recovered studies is consistent with the relatively sparse literature on other early childhood self-regulation programs. For example, the Chicago School Readiness Project, which, like Tools, is a new school-based self-regulation

intervention, has only one RCT-based evaluation study (Raver et al., 2011), whereas Montessori, a well-established early childhood curriculum, has no RCT studies in its evidence base (Institute of Educational Sciences, 2016; Lillard, 2005).

Similarly, the Incredible Years, described in Chapter Six, has been implemented for over thirty years in more than twenty countries (K. Jones et al., 2007), yet the number of existing studies for any one of its multiple intervention arms is notably low. For example, Nye (2013) analyzed the evidence base for the Incredible Years Teacher Classroom Management program (TCM), which, like Tools, is a school-based intervention. Nye (2013) recovered only four studies to include in its multilevel meta-analysis, even though Incredible Years has been in existence for longer than Tools.

The most recent systematic review, to the best of my knowledge, of early childhood executive function interventions (Jacob & Parkinson, 2015) also reported difficulties with identifying a large number of evaluation studies. In fact, the authors reported outcomes from only one study on both the Chicago School Readiness Project as well as the Head Start REDI program (see Chapter Six). Thus, the small number of included studies in the present review mostly mirrors the limited evidence observed for other similar programs.

Some may wonder whether the present study simply missed other existing Tools studies; however, the exhaustiveness of the search strategy would suggest that this was not the case, especially considering that this project occurred over a three-year timespan. In fact, this research synthesis project uncovered both the original studies and raw data for multiple published and unpublished studies, which is an issue toward which we now turn.

10.1.2 Potential publication bias in the existing Tools literature base

Across the seven recovered Tools studies in this analysis, the only previously published studies (Barnett et al., 2008; Blair & Raver, 2014; Diamond et al., 2007) all indicate statistically significant benefits for children in the Tools program. By contrast, the studies that indicate null (Clements & Sarama, 2012; Lonigan & Phillips, 2012; Morris et al., 2014) or negative (Farran & Wilson, 2014) results have not been published. Thus, the published

research may overestimate Tools' effect compared to the full body of evidence (i.e., both the published and unpublished studies).

Although a minimum of ten studies is required to conduct formal publication bias assessment via funnel plot or moderation analysis, the contrast between the published and unpublished Tools results emerges as preliminary evidence of publication bias. Many educators and policymakers may be relying on the positive findings from the published studies when choosing to implement Tools, yet they may not be aware of the null and negative findings in the unpublished Tools literature.

To the best of my knowledge, this potential publication bias has not been mentioned in either the academic literature or policy debates about Tools. In fact, the Tools website, which has a section dedicated to research about the program, includes only the published literature even though the Tools developers worked directly with Farran & Wilson (2014), who found negative self-regulatory effects for Tools students over time.

In order to rigorously assess publication bias, the present review will be updated every three years with new research as dictated by my Campbell Collaboration protocol. Thus, the current finding of potential publication bias, if corroborated through future updates to this review, would represent another substantial contribution to the literature. The present review's other main contribution to the literature is the first meta-analysis of Tools' quantitative effects, the results of which are now discussed below.

10.1.3 Summary of the meta-analytic results

The meta-analytic results in section 9.7 indicate positive effect sizes that favor Tools across the four outcome measures. However, those effect sizes did not reach statistical significance for Tools vis-à-vis comparator curricula across three outcome measures: 1) assessor report-based ratings of children's self-regulation, 2) task-based self-regulation indicators, and 3) literacy skills. By contrast, statistically significant impacts were observed for the math composite effect size among Tools students.

The significant math effects for Tools is especially noteworthy given that the Tools developers consider math to be an area of weakness for the curriculum (Mackay, 2013). One

possible explanation for the observed effect is that few early childhood programs allocate time to math at all (Weiland & Yoshikawa, 2013); thus, even though Tools may not have an especially strong math regimen, Tools students may still have been exposed to more math than students in many comparator classrooms. Because this study lacks data on the time spent in math among comparator classrooms, this hypothesis cannot be rigorously tested.

On the topic of hypotheses, it is important to note the lack of formal hypotheses presented for this first study. In fact, given that this is a research synthesis project, it is not common to have research hypotheses, which could introduce bias into the research process. That is, by explicitly stating expectations about the research outcome, it is possible for researchers to introduce bias in the way they analyze data (i.e., researchers might, consciously or unconsciously, analyze data in such a way that the results align with their expectations).

Thus, although I did not incorporate hypotheses into this research synthesis study, the Tools developers have repeatedly hypothesized (Bodrova & Leong, 2007, 2013; Leong & Bodrova, 2011) gains for Tools students, especially for self-regulation. In the results presented here, the effect sizes were all in the positive direction for Tools students, but the effect sizes were small (i.e., $\max = .17$) and statistically insignificant for both self-regulation outcomes. Consequently, despite potentially promising evidence from the positive effect sizes in favor of Tools, more research is necessary to demonstrate that those effects are statistically significant as opposed to arising from chance alone.

Although the null statistical effects did not align with the developers' expectations, the results do remain consistent with many outcome evaluations of early childhood programs. For example, a national evaluation of 14 preschool curricula in the United States (Preschool Curriculum Evaluation Research Consortium, 2008) found that none of the curricula significantly improved children's self-regulation skills beyond comparator curricula.

Moreover, Chapter Seven of this dissertation outlined the paucity of evidence for self-regulation promotion among multiple early childhood curricula. That is, none of the curricula reviewed in Chapter Seven had rigorous evidence for improving self-regulation. This is not to say that children's self-regulation did not improve in any of the curricula; rather, the studies showed no evidence that one curriculum promoted children's self-

regulation significantly more than any of the other sampled curricula. Thus, the absence of an observed effect for Tools perhaps may not be surprising, even though Tools explicitly claims to hone children's self-regulation skills.

Moreover, it is critical to distinguish between the notions of 'no evidence of effect' versus 'evidence of no effect.' As Littell et al. (2008) explain, "no evidence of an effect is *not* the same as evidence of no effect; insufficient statistical power (too few studies, too much heterogeneity) is an alternative explanation for null results" (135). In other words, we cannot conclude that Tools *does not work*; rather, the evidence produced here simply does not conclusively demonstrate that Tools *does work* as designed.

The reasons for the absence of an observed effect may be related to factors external to the program itself, such as the statistical power and heterogeneity issues to which Littell et al. (2008) refer. Those two issues, as well as other related ones, are addressed in the limitations section (10.5).

10.2 Quality of the evidence base

Many evaluative protocols exist to assess the quality of evidence bases in systematic reviews. The framework used most widely by the Cochrane Collaboration is called Grades of Recommendation, Assessment, Development, and Evaluation (GRADE; Guyatt, G. H., Oxman, A. D., Schünemann, H. J., Tugwell, P., & Knottnerus, 2011). In the GRADE framework, evidence bases comprised of randomized controlled trial studies receive the highest possible rating (i.e., high-quality, medium quality, low-quality, and very low quality). A 'high-quality' rating implies that the estimated effect in the meta-analysis closely resembles the true program effect in the population (Guyatt, et al., 2011).

To evaluate the Tools evidence base, then, it is important to note that the present systematic review contains evidence from seven randomized controlled trial studies. Thus, the systematic review contains no quasi-experimental or observational studies, which provides initial confidence in the quality of the existing evidence base. However, two of the included studies exhibited a high risk of attrition bias because the authors did not analyze differences between the attrited and non-attrited participants. Moreover, the fact that only seven studies

were retrieved led to issues of low power, which the GRADE framework also identifies as a problem for meta-analysis (Guyatt, et al., 2011).

Thus, although the overall Tools evidence base begins with a high-quality rating for including only randomized controlled trials studies, the risk of bias and low power issues result in a downgrade of the quality rating. Unfortunately, the GRADE guidelines do not clearly distinguish between a moderate versus low quality rating. According to the GRADE guidelines (Balshem, H., Helfand, M., Schünemann, H. J., Oxman, A. D., Kunz, R., Brozek, J., & Guyatt, 2011), a ‘moderate’ rating suggests that “the true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different” (p. 404), whereas a low rating indicates that “the true effect may be substantially different from the estimate of the effect” (p. 404).

Given the lack of clarity surrounding moderate versus low ratings in the GRADE framework, it is difficult to decisively state whether the Tools evidence base is of moderate or low quality. This is because we do not, and cannot, know whether the true effect is different from our estimate of the effect, which complicates the determination of a moderate versus low quality GRADE rating.

Nonetheless, the seven studies included here were of mostly high quality as individual studies (i.e., only two of the studies exhibited high-risk of bias across the six Cochrane risk of bias dimensions). Thus, I argue that the Tools evidence base is of moderate quality, though others could reasonably argue for a low quality rating as well. In the future, it is hoped that more clear GRADE guidelines will be developed to differentiate between moderate and low quality evidence bases.

10.3 Differences between protocol and review

There are no differences between the research plan I had outlined in my online protocol and this review. The only missing element of the final review is the moderation analysis. That is, the protocol indicated a plan to conduct various moderation analyses, but the protocol does state that such analyses would only be conducted if a sufficient number of studies (i.e., ten per moderator) were recovered. Since the requisite number of studies was not recovered,

this review adhered to the protocol's directives for moderation analyses as well as the other components of the systematic review protocol.

10.4 Strengths

This study has two noteworthy strengths. The first is the sophistication of the statistical approach, and the second is the exhaustiveness of the systematic search. Each strength is briefly discussed in turn below.

10.4.1 Sophistication of the statistical approach

According to Littell et al. (2008), meta-analysis has become “more widely used in the social sciences, especially in psychology and education” (p. 4). Although meta-analysis has become more widespread, traditional meta-analytic techniques do not allow for the incorporation of multiple related effect sizes from the same study. Thus, the present study used multilevel meta-analysis, which modern statistical software has made executable only in recent years (Viechtbauer, 2010).

Many traditional meta-analyses simply average all the effect sizes within studies before computing the composite effect size, but this leads to a loss of data. For example, if I had averaged the effect sizes for the assessor-reported and researcher-reported self-regulation skills within each study, then I would have had one effect size instead of two, which reduces statistical power and increases Type II errors (i.e., failing to detect a statistically significant effect that truly exists in the population).

By contrast, this review incorporates all of the originally reported relevant effect sizes in the multilevel meta-analysis (as well as a robustness check with robust variance estimation), thereby retaining as much information as possible while accounting for the statistical dependency among effect sizes nested within the same study.

10.4.2 Exhaustiveness of the meta-analysis data collection process

Although the present research synthesis contains only seven studies, I am confident that the search captured every available study on the Tools program. Originally, only two of the

seven included studies published the requisite data for me to execute a meta-analysis. I obtained the remaining data through emails and calls to the other study authors, who took between 24 hours and two years to release their data.

In the end, six of the seven research teams in question opted to provide their raw data to incorporate in this meta-analysis. The only research team to refuse data release (Lonigan & Phillips, 2012) was still included into the narrative summary portion of the systematic review while excluded from the quantitative synthesis.

Moreover, and more generally, the present dissertation has been in process for three years, which has provided ample time to identify every extant Tools research program. Thus, the results presented here, to the best of my knowledge, represent the most comprehensive analysis of the Tools evidence base to date.

10.5 Limitations

Despite the strengths described above, this study has multiple limitations, three of which are described below: 1) the small number of included studies, 2) the variation in Tools implementation across studies, and 3) the potential for measurement error across studies.

10.5.1 Small number of included studies

Although the literature was systematically and exhaustively searched for this study, only seven studies fit the methodological criteria outlined in Chapter Eight. A small number of studies is not necessarily a problem in and of itself. For example, in *Practical meta-analysis*, Lipsey and Wilson (2001) explain that meta-analysis can be responsibly conducted with “as few as two studies” (7), which has been corroborated by other authors (Hox, 2002; Valentine, Pigott, & Rothstein, 2010).

The application of this logic to the multilevel meta-analysis context is less clear, however. Dr. Josh Polanin, one of the leading research synthesis methodologists at the Campbell Collaboration, has noted (Polanin, 2013) that there is no clear minimum number of studies to be included in a multilevel meta-analysis. According to Tanner-Smith and Tipton (2014), a small number of studies, which they define as fewer than ten, is especially problematic

when researchers attempt to conduct moderation analyses through meta-regression. Given that this study did not conduct moderation analysis, and given the consistency of the effect size estimates across the multilevel and robust variance estimation frameworks, it is likely that the small number of included studies was not a problem in this study. This is especially true given the large number of effect sizes per included study, which lends additional power to the analyses (Hox, 2002; Van den Noortgate, López-López, Marín-Martínez, & Sánchez-Meca, 2014).

In fact, in my literature search, the first methodology paper found to apply multilevel meta-analysis to educational data (Goldstein, Yang, Omar, Turner, & Thompson, 2000) contained only eight randomized control trial studies. Despite the small sample size, the authors found significant effects for class size, where larger class sizes predicted lower literacy achievement. The authors of that multilevel meta-analysis (Goldstein et al., 2000) and others (Hox, 2002; Tanner-Smith & Tipton, 2014) have noted that statistical power is heightened with more studies, but, once again, multilevel meta-analysis can still be responsibly conducted with a relatively small number of studies.

Moreover, the alternative to a multilevel approach in traditional meta-analysis is either to discard several effect sizes from a study or to average all effect sizes within a study. Both of those approaches result in a substantial loss of information because the effect sizes are either aggregated into an often poorly defined composite (e.g., several task-based and informant-reported self-regulation measures bundled into one composite with measurement error) or abandoned altogether. By contrast, the multilevel approach used here incorporated all available effect sizes to capture a maximally comprehensive view of Tools' associations with child outcomes.

That said, the small number of studies included in this Tools meta-analysis still creates two central problems: 1) the analyses may be insufficiently powered to identify statistically significant results, and 2) the small study sample size precludes moderation analyses. On the first point, the notion of statistical power (i.e., the minimum sample size necessary to detect a significant effect in the sample if it truly exists in the population) applies both to primary research methods as well as secondary research methods such as meta-analysis. If a meta-

analysis has a small number of studies (i.e., fewer than ten), then “inadequate statistical power is a possible explanation for null results” (Littell et al., 2008, p. 130).

Once again, all four outcome measures had effect sizes that favored the Tools condition, but three of the four were statistically insignificant. It is possible that the observed null results in the present study may have reached statistical significance with more studies in the sample, which will become clear as I update this review in the future (see section 10.6.1). In the meantime, however, it is important to note that statistically significant results were observed for math, even though the composite math effect size had only 15 constituent effect sizes across three studies included in the analysis.

By contrast, literacy and task-based self-regulation each had more than 30 effect sizes nested within five studies each, but neither of those composite effect sizes achieved statistical significance despite having higher statistical power. Thus, although power may have played a role in the null effects, it is certainly not the only explanatory factor in the observed results given that the math effect size achieved significance with comparatively lower power.

The second problem, regarding moderation analyses, is inextricably linked to the first problem of low power. That is, without at least ten studies per moderator, there is not enough power to run moderation analyses. In the present study, only six studies were ultimately included in the quantitative synthesis. Once again, I will incorporate moderation analysis as I update this review in the future.

10.5.2 Variation in Tools implementation across studies

As described in section 9.4, the included studies did not implement Tools in the same way. Some studies compared Tools on its own versus a comparison curriculum, whereas other studies combined Tools with another intervention (e.g., the Building Blocks math curriculum in Clements et al., 2014) to compare against a control group. Thus, while the meta-analytic results aim to assess Tools’ effect vis-à-vis comparator curricula, neither the Tools nor comparison curricula can be considered as a monolith. Specifically, the comparison group curricula varied across all studies, and each Tools study also probably implemented Tools slightly differently.

This is especially likely given that the included studies were all conducted in an American context, where early childhood education standards and approaches have been shown (Gilliam, W. S., & Zigler, 2004) to vary substantially across the fifty states. In this research synthesis, the included reports did not extensively describe their 'Tools' implementation approach, so we cannot know how Tools manifested differently in each study across different contexts.

For example, one of the included studies (Clements et al., 2014) modified Tools by combining it with the Building Blocks math curriculum; previous research (Elliott, D. S., & Mihalic, 2004; Mihalic, S., Fagan, A. A., Irwin, K., & Ballard, 2002) has indicated that programmatic adaptations can sometimes omit core components of interventions, which may alter the program's observed efficacy.

Nevertheless, the present study did conduct sensitivity analyses to determine whether 'Tools' effect varied across situations where Tools was implemented as a stand-alone curriculum versus situations when Tools was implemented alongside another program. That sensitivity analysis (see section 9.10) indicated no significant differences between Tools implementation as a stand-alone versus as a combined program. Nonetheless, the inherent Tools implementation variation as well as the variability among comparator curricula across included studies introduces uncertainty into the results.

10.5.3 Potential for measurement error across studies

Given that the meta-analysis revealed null effects across three of the four outcome measures, it may seem sensible to conclude that Tools has no effect on children's self-regulatory and literacy skills. Once again, however, no evidence of effect is not the same as evidence of no effect (Littell et al., 2008). Instead, it is possible that other factors masked the impact of the Tools program on child outcomes.

One such factor could be measurement error in the assessment instruments across studies. Although the included studies exclusively employed standardized testing instruments, it remains possible that those instruments, especially for self-regulation, had low construct

validity, which has been noted in the self-regulation measurement literature (McClelland & Cameron, 2012).

Thus, it is possible that low reliability and validity across the measures may have contributed to the observed null results. As Kline (2015) writes, measurement error “generally reduces effect sizes below their true (population) values” (p. 92). Consequently, Kline (2015) recommends the use of latent variable models, which account for measurement error in order to more accurately capture relationships among phenomena (e.g., between the Tools program and self-regulation).

Whereas the upcoming second study (Section IV) will use latent variable models to account for measurement error in its analyses, none of the studies included in the present meta-analysis used latent variable models. Thus, it is possible that measurement error attenuated the true relationship between Tools and children’s outcomes. It is also possible that measurement error did not have such an effect, but this hypothesis cannot be tested without the raw data from all included studies. Fortunately, the upcoming Section IV analyses use raw data from one of the included studies (Farran & Wilson, 2014) to investigate the relationship between Tools activities and children’s self-regulation.

10.6 Suggestions for future research

Several possibilities for future research would strengthen the existing literature base. Once again, the measurement error limitation described in section 10.5.3 will be addressed in the upcoming Section III analyses, so that research program will not be further described here. Instead, three additional directions for future research are now described.

Specifically, three possibilities for future research include: 1) continued updates to the present meta-analysis, 2) a multi-arm cluster trial that directly compares Tools with other self-regulation interventions and curricula, and 3) a meta-analysis of several early childhood interventions and curricula. Each is described in turn below.

10.6.1 Continued updates to the present meta-analysis

Given that Diamond et al. published the first Tools evaluation study only in 2007, and given the ongoing expansion of the Tools program, it seems likely that Tools research will similarly expand in the future. As new studies are released, it is critical to update the current meta-analysis with new data.

In fact, section 9.3 already detailed one Tools study that remains ongoing, which is the first Tools evaluation study to specifically investigate Tools with English language learners. Moreover, an email correspondence with an expert in the field, Dr. Behrman, (personal communication, April 10, 2016) suggested that a Tools evaluation study from Chile will be published in the coming years, though I was unable to gather more specific information about that study.

Consequently, as more Tools studies are published, it will become possible to conduct moderation analysis to see how the curricular effect varies by child-level characteristics (such as children's language learning status) as well as study-level characteristics (such as whether Tools was implemented in South versus North America).

Once again, my Campbell Collaboration protocol (Baron et al., 2016) explicitly requires me to update the Tools meta-analysis every three years. Thus, as more Tools evaluation studies are published over time, I will be able to extend upon the current project through the inclusion of moderation analysis, publication bias assessment, and potentially other meta-analytic methodological innovations that remain in development.

10.6.2 A multi-arm trial comparing Tools to other self-regulation programs

Five of the seven included Tools evaluation studies compared Tools against a single control or 'business-as-usual' condition. By contrast, the other two studies compared Tools against both a 'business-as-usual' group as well as another intervention group: Lonigan & Phillips (2012) used the Literacy Express curriculum and Clements et al. (2014) used the Building Blocks math curriculum. These latter two studies enable assessment of the relative effectiveness among Tools, another target early childhood curriculum, and a 'business-as-usual' program.

Unfortunately, in the existing research, the two multi-arm trials compared Tools against literacy and math curricula, respectively, instead of against the other self-regulation interventions described in Chapter Six. In the future, it is hoped that researchers will conduct a large-scale, multi-arm trial that compares Tools to other self-regulation interventions such as Incredible Years, the Chicago School Readiness Project, the Promoting Alternative Thinking Strategies program, and other programs all within one randomized trial. In so doing, educators and researchers can directly compare Tools' effectiveness against other programs with similar aims.

In the present meta-analysis, we observe that Tools did not predict significantly improved task-based or assessor-reported self-regulation relative to the set of comparator curricula. However, the set of comparator curricula largely involve programs such as HighScope, Creative Curriculum, and others. Like Tools, those are all comprehensive curricula, but, in contrast to Tools, they lack a specific focus on self-regulation. Thus, a multi-arm trial comparing Tools with several other early childhood self-regulatory interventions could represent a significant contribution to the literature.¹³

10.6.3 A meta-analysis of several early childhood interventions

Whereas the research plan described in the previous section refers to a single, large-scale trial that concurrently analyzes multiple self-regulation interventions, another research program could conduct a large-scale meta-analysis that aggregates data across multiple studies regarding multiple interventions. Many meta-analyses do analyze multiple interventions, but, to the best of my knowledge, no meta-analyses have investigated multiple self-regulation interventions' impact on children's self-regulation skills.

One recent systematic review (Jacob & Parkinson, 2015) did investigate the impacts of school-based executive function interventions on children's academic achievement. However, that study did not include any Tools research in its meta-analysis, and it also used only academic measures as outcomes instead of self-regulation measures. A future review

¹³ A potential model for this future study could be the 2008 United States government-funded Preschool Curriculum Evaluation Research Consortium (PCER) study, which concurrently analyzed 14 curricula across the country (Preschool Curriculum Evaluation Research Consortium, 2008) using a multi-arm, cluster randomized controlled trial design.

could expand on Jacob and Parkinson's (2015) study by investigating several self-regulation interventions' effects on children's self-regulation outcomes instead of only academic skills.

10.7 Chapter summary

This chapter discussed the results for study one, which stands as the first Tools systematic review and meta-analysis. Given the future research possibilities explained in section 10.6, it is hoped that the findings of this review will be replicated and extended in other studies, including through my own updates to the Campbell Collaboration review.

Moreover, study two of this dissertation, toward which we now transition, explicitly aims to extend upon the meta-analytic findings described here. Specifically, study two is the first to analyze the associations between individual Tools learning activities and children's self-regulation skills. After the methods, results, and discussion for study two have been shared in the upcoming Section IV, the findings from studies one and two are briefly synthesized in Section V. The fifth and final section also describes these findings' implications for policy and practice while bringing the dissertation to a close.

SECTION IV: METHODOLOGY, RESULTS AND DISCUSSION FOR STUDY TWO

CHAPTER 11: Methodology for study two

This chapter outlines the methodology for study two, which concerns Tools of the Mind's (Tools) overall effect on self-regulation as well as the associations between Tools activities and self-regulation. Study two aims both to replicate a previous Tools evaluation study (Farran & Wilson, 2014) and to be the first analysis of specific Tools activities. In so doing, study two aims to identify instructional practices that can be replicated within and beyond the Tools context to improve children's self-regulation.

To address this aim, the present study employs multilevel structural equation modeling techniques to analyze data from 1145 children in 80 U.S. classrooms. This chapter first describes the data sample, measures, and ethics for study two. Next, this chapter presents the research questions and hypotheses for study two. Finally, this chapter reviews the statistical methods for study two.

11.1 Data sample

Study two relies upon data from Vanderbilt University's Peabody Research Institute (PRI). PRI earned a grant from the U.S. Department of Education to conduct a cluster-randomized evaluation of Tools' effectiveness (Farran & Wilson, 2014). Consequently, PRI collected data on 1145 children in 80 American preschool classrooms in 59 schools across six school districts in two states. The data include an executive function test battery, teacher and researcher reports of children's self-regulation, as well as teachers' implementation frequency and implementation fidelity of Tools activities. Thus, these data are suitable for the present study to analyze the association between Tools activities and student outcomes.

11.2 Design and procedures of the PRI study

To conduct the experimental evaluation of Tools, PRI researchers recruited schools in two states through a two-cohort structure. Cohort one had 60 classrooms in 47 schools that were randomly assigned to either Tools or 'business-as-usual' (Farran & Wilson, 2014, p. 9) comparison curricula, whereas cohort two had 20 classrooms in 12 schools.

Cohort one classrooms that were randomly assigned to Tools (n=32) began Tools implementation in 2009, whereas cohort two classrooms randomly assigned to Tools (n=10) began implementation in 2010. Farran & Wilson (2014) selected a two-cohort model to evaluate reliability; that is, would the results be consistent across two cohorts given identical experimental designs? This study includes data from both cohorts to address this question.

The child-level demographic data for each cohort are presented below in table 9. The second cohort had significantly more ($\chi^2_{[1]} = 24.56, p < .01$) English language learner students than the first cohort. Beyond English language learners, however, no other significant differences emerged between Tools and control students across the two cohorts. Differences in results across the two cohorts are presented in section 12.2.5.

Table 9: Cohort 1 and 2 differences between Tools and control classrooms

	Tools condition		Control condition	
	n	%	n	%
Cohort 1				
Male	261	52.4	218	57.5
White	192	38.6	157	41.4
Black	145	29.1	86	22.7
Hispanic	118	23.7	95	25.1
Asian	32	6.4	21	5.5
Multi-racial	4	0.8	16	4.2
Other minority	7	1.4	4	1.1
English language learners (yes)	140	28.1	117	30.9
Special education plan (yes)	68	13.7	58	15.3
Cohort 2				
Male	80	54.4	64	53.8
White	56	38.1	62	52.1
Black	45	30.6	24	20.2
Hispanic	39	26.5	25	21
Asian	2	1.4	4	3.4
Multi-racial	4	2.7	2	1.7
Other minority	1	0.7	2	1.7
English language learners (yes)	60	40.8	62	52.1
Special education plan (yes)	14	9.5	6	5

The project's chronological progression is depicted in figure 16. First, teachers in the experimental group received Tools training for one academic year (2009-2010). No data were collected during that time to ensure that the Tools group educators became familiar with the program. During the following school year (2010-2011), PRI researchers began conducting child assessments and observations in each classroom. 2010-2011 was the only school year during which Tools was implemented, and the children had their first assessment battery during autumn 2010 of that pre-kindergarten school year.

The assessment batteries were re-administered in the spring of students' pre-kindergarten, kindergarten, and first grade years. The average age of the children at each time point was 54 months ($SD = 3.63$) during pre-kindergarten autumn, 62 months ($SD = 3.56$) during pre-kindergarten spring, 73 months ($SD = 3.62$) during kindergarten spring, and 85 months ($SD = 3.62$) during first grade spring. Thus, on average, children began Tools at approximately four-and-a-half years old, completed it after their fifth birthday, and then were followed up with assessment batteries at ages six and seven during kindergarten and first grade, respectively, to assess whether the intervention effect was sustained past implementation.

In contrast to the child assessment batteries, the classroom observations occurred only in autumn, winter, and spring of the pre-kindergarten year. Thus, the observations were not repeated past pre-kindergarten, which was the only implementation year of Tools (Wilson & Farran, 2012). After the pre-kindergarten year, all students moved to kindergarten classrooms, none of which implemented Tools (Wilson & Farran, 2012). Thus, although this study collects data over three years, the study was designed such that Tools was only implemented and observed during the first of the three years.

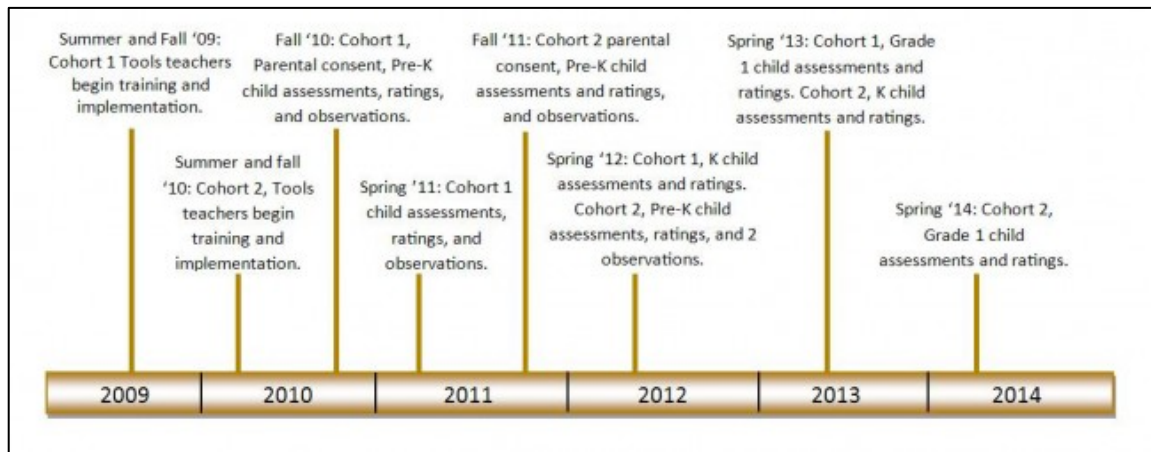


Figure 16: Chronological progression of the Peabody Research Institute's (PRI) Tools study

11.2.1 School recruitment

In cohort one, 646 children in 60 classrooms in 47 schools agreed to participate. In cohort two, 499 students in 20 classrooms in 12 schools agreed to participate. Rural, suburban, and urban settings were all represented among the six total districts. The names of the participating schools, teachers, and students were removed from the data to ensure confidentiality (Farran & Wilson, 2014).

Once again, PRI employed a cluster-randomized design to test 'Tools' effectiveness versus the pre-existing curricula used in the comparison group schools. Clustered designs are standard practice (Bryman, 2012; Punch, 2009) for data that are naturally nested (i.e., micro-level units nested within increasingly macro-level units). Given that the PRI data involve students nested within classrooms nested within schools, a clustered design was necessary.

Specifically, schools were selected as the cluster unit of randomization. That is, the random assignment procedure dictated which schools (and all teachers and students within them) engaged with Tools. Selecting schools as the cluster unit of randomization minimized the likelihood that Tools and non-Tools teachers would share instructional practices that could have "compromised the experimental contrast" (Farran & Wilson, 2014, p. 10). Schools assigned to the comparison condition continued to use their pre-existing curricula, which primarily consisted of Creative Curriculum, Opening the World of Learning, and Building Blocks (Fuhs et al., 2015, p. 210).

11.2.2 Assessment procedures

PRI collected child-level data between 4 – 6 weeks after the start of the pre-kindergarten school year and between 4 – 6 weeks from the end of the pre-kindergarten year (Fuhs et al., 2015, p. 212). To test whether Tools affected children's self-regulation past the pre-kindergarten implementation year, PRI used the same measures in the spring of both the kindergarten and first grade year.

PRI researchers were trained to conduct the assessments, which lasted approximately 20 minutes per child (Farran & Wilson, 2014). The sessions were conducted with a fixed sequence of assessments at each time point.¹⁴ The average interval between autumn and spring testing administrations was 7.38 months with a standard deviation of .55 months.

11.2.3 Classroom observation procedures

During the Tools pre-kindergarten implementation year, PRI conducted full-day classroom observations during the autumn, winter, and spring, resulting in three total observations. The observations were used to collect data regarding teaching practices, learning activities, and Tools implementation fidelity (see section 11.2.4) below.

To confirm inter-rater reliability, estimates were computed for whether a specific Tools activity was occurring (e.g., buddy reading, make-believe play, or other activities), and whether a Tools mediator was being used (e.g., buddy reading role cards to help each child identify and adhere to his or her role). The Cohen's kappa reliability estimates were .95 and .90, respectively, which are above both Kline's benchmark of .70 (Kline, 2015) and Bryman's of .80 (Bryman, 2012).

11.2.4 Implementation fidelity

During the classroom observation periods described in section 11.2.3 above, PRI and the Tools curricular developers also assessed each teacher's execution of Tools using an

¹⁴ In the first executive function testing session, children completed the Peg Tapping, Hands-Toes-Knees-Shoulders, and Copy Design tasks. The second testing session included the Dimensional Card Change Sort and Corsi Blocks tasks (Fuhs et al., 2015, p. 212). See section 11.5.1 for descriptions of each test.

implementation fidelity measure (Vorhaus & Meador, 2010), which was created for the purposes of the PRI study (Farran & Wilson, 2014).

Fidelity evaluation enabled assessment of whether child outcomes resulted from Tools' effectiveness or from different levels of Tools implementation fidelity. That is, if Tools students achieved low self-regulation scores but experienced poor implementation of Tools, then it would be unclear whether students' results should be attributed to the teacher's poor Tools implementation or to Tools' ineffectiveness. Thus, PRI's fidelity measure enables more accurate estimation of Tools' impact on child outcomes.

Overall, the PRI researchers reported high levels of Tools implementation fidelity across teachers in the sample, with the minor variation in fidelity not associated with children's self-regulation outcomes (Farran & Wilson, 2014, p. 19). Nevertheless, teachers' fidelity of implementation will be further tested in the present study as described in section 11.7.2.

11.3 Analytic approach: Secondary data analysis

Because the present study uses existing data but addresses new research questions, this study's approach is characterized as secondary data analysis. Secondary data analysis is formally defined as "using pre-existing data in a different way or to answer different research questions than intended by those who collected the data" (Schutt, 2011, p. 306).

Secondary data analysis provides the researcher with access to high-quality data without the burden of personally collecting the data. As such, secondary data analysis fulfills the needs of a researcher with "macro-interests but micro-resources" (Glaser, 1963, p. 11). Because datasets such as PRI's are collected on a large scale, the sample sizes are sufficiently large to detect true effects within a representative sample, which is the appropriate data type with which to inform policy (Smith, 2008). Overall, the access to high-quality data and the ability to make inferences based on large samples are the key benefits of secondary data analysis.

The oft-cited disadvantage of secondary data analysis is that the secondary researcher is not present for the primary data collection (Punch, 2014). As such, the secondary researcher is beholden to the quality of the data provided. Although this disadvantage is problematic, it is linked to the first advantage described above. Specifically, secondary datasets are often

much larger and more comprehensive than researchers such as myself have the capacity to independently collect. Moreover, given my personal contact with the PRI researchers who collected the data (see the Ethics section below), I was able to clarify any remaining questions about the data throughout the research process.

11.4 Ethics

Even though this dissertation uses pre-existing anonymous data, it remains necessary for the secondary researcher to gain ethical approval (Howe & Moses, 1999). Thus, before beginning the study, ethical approval was obtained through the Departmental Research Ethics Committee. The ethical approval notification is attached in Appendix A.

After receiving ethical approval, I requested the PRI dataset during an in-person meeting with researchers at Vanderbilt University. The first meeting took place on August 5, 2014, during which the PRI team agreed to share the dataset with me (e-mail text attached in Appendix F). Once the data download was complete, all datasets were saved in their original anonymous format on two encrypted drives.

All participating schools, teachers and parents provided opt-in consent to researchers at the beginning of the PRI study; moreover, all children provided verbal assent prior to each testing session that yielded the child-level self-regulation data (Farran & Wilson, 2014), which are described below in the measures section.

11.5 Self-regulation measures

As with the meta-analysis in study one, the analyses in study two also distinguish between task-based and informant-report self-regulation measures. Task-based measure scores derive from children's performance on executive function tasks that assess cognitive control; by contrast, informant-report measures quantify children's self-regulation according to ratings from researchers or teachers. The sections below describe the specific measures associated with each of the two categories in more detail.

11.5.1 Task-based self-regulation measures

Once again, the term ‘task-based’ signifies that children’s self-regulation capacity is being directly measured according to their score on a task. As such, the assessor is not providing a rating or observational measure of the child’s self-regulation. For each of the following measures, children receive directions to perform a task, and a researcher records their score.

Corsi Blocks

The Corsi Blocks task (Corsi, 1972) requires children to tap a set of blocks in the reverse order from that demonstrated by an examiner. The task begins with two blocks and then increases by one block in each successive trial. The assessment stops after a child responds incorrectly for two consecutive trials on any given pattern length. A child’s score represents the longest backward pattern completed by the child (Corsi, 1972).

This tests all three aspects of executive function: working memory to remember the pattern, cognitive flexibility to reverse the pattern demonstrated by the examiner, and inhibitory control to suppress the impulse to imitate the examiner’s pattern. In children between the ages of 4 and 11, the Corsi Blocks task has strong ($r = .83$) test-retest reliability (Alloway, Gathercole, & Pickering, 2006).

The unstandardized means, standard deviations, skewness, and kurtosis values for the four assessment time points are listed below in table 10. The values in the tables below derive from imputed data (see section 11.9); thus, data from all 1140 children in the analytic sample were used to compute the descriptive statistics for each of the outcome measures. The values in table 10 indicate mean score increases and stable dispersion across time. Moreover, the data are normally distributed, as assessed by Kline’s (2015) guidelines of skewness (all values are less than $|2|$) and kurtosis (all values are less than $|20|$).

Table 10: Corsi Blocks descriptive statistics at each time point

Item	Mean	SD	Skewness	Kurtosis
CB_t1	1.14	1.13	0.24	-1.25
CB_t2	1.55	1.32	0.16	-0.94
CB_t3	2.82	1.36	-0.48	-0.06
CB_t4	3.73	1.27	-0.80	0.84

Dimensional Change Card Sort task

The Dimensional Change Card Sort task (DCCS; Zelazo, 2006) requires children to sort a set of cards across one dimension (e.g., color), and then to sort the set of cards across another dimension (e.g., shape). In the PRI study (Farran & Wilson, 2014), children had to first sort cards into red and blue piles, and then sort the cards between star and truck shapes.

If children successfully sorted cards across those two dimensions, then a third dimension was added: whether or not the card had a black border. If the card had no border, then children had to sort cards by truck versus star shape. If the cards did have a border, then children were instructed to sort the cards by their color. Finally, if the child correctly executed that trial, then the rules switched, and the child had to complete this fourth permutation (i.e., no border indicates sort by color and vice versa).

The scoring procedures contain five possible outcomes. Firstly, an unsuccessful initial sort by both color and then shape results in a score of 0. Secondly, children who pass the initial sort by color but not by shape receive a score of 1. Thirdly, children who pass both the shape and color but not the border sort receive a score of 2. Next, children who pass the color, shape, and border sort receive a score of 3. Finally, children who pass a reversed trial where the border contingency rule is reversed receive a score of 4.

In children between ages three and seven, DCCS exhibits strong ($ICC = .92$) test-retest reliability and is a standardized measure in the National Institutes of Health (NIH) Toolbox Cognition Battery (Zelazo et al., 2013).

The unstandardized means, standard deviations, skewness, and kurtosis values for the four assessment time points are listed below in table 11. The means increase over time, and the standard deviations are consistent until the fourth time point, when the dispersion across children increases and some children earn the score of 4. The skewness and kurtosis values are all within acceptable levels.

Table 11: Dimensional Change Card Sort descriptive statistics across time points

Item	Mean	SD	Skewness	Kurtosis
DCCS_t1	1.29	0.58	0.14	-0.14
DCCS_t2	1.65	0.59	-0.35	-0.02
DCCS_t3	2.01	0.60	-0.40	1.52
DCCS_t4	2.53	0.92	0.44	-0.69

Peg Tapping

The Peg Tapping task (Luria, 1966) requires children to tap a peg a certain number of times based on rules dictated by the examiner. Specifically, a child must tap twice when the examiner taps once or once when the examiner taps twice. In this task, children receive two practice trials with feedback followed by 16 test trials without feedback. Children receive a score of 0 for incorrect responses and 1 for correct responses. Final scores thus range between 0 and 16.

This task taxes all three elements of executive function: working memory to remember and act on the instructions, cognitive flexibility to switch among the tapping patterns, and inhibitory control to suppress the impulse to imitate the examiner's tapping pattern. The Peg Tapping task exhibits high ($r = .74$) test-retest reliability (Nampijja et al., 2010).

The unstandardized means, standard deviations, skewness, and kurtosis values for the four assessment time points are listed below in table 12. The means increase over time, whereas the dispersion decreases, which indicates that children's Peg Tapping scores become more similar over time. Finally, the skewness and kurtosis values are acceptable at time points one, two, and three but unacceptable at time point four. The problems with skewness and kurtosis were treated through a latent variable transformation (see section 11.6).

Table 12: Peg Tapping descriptive statistics across time points

Item	Mean	SD	Skewness	Kurtosis
PT_t1	4.62	5.40	0.77	-0.88
PT_t2	9.28	5.58	-0.50	-1.18
PT_t3	13.30	3.90	-2.01	3.53
PT_t4	14.75	2.37	-3.64	16.59

Heads-Toes-Knees-Shoulders

Head-Toes-Knees-Shoulders (HTKS; Ponitz et al., 2009) requires children to listen to an experimenter’s instruction and then perform the reverse bodily movement on themselves. Specifically, children receive two oral prompts: “touch your toes” and “touch your head.” Children must perform the opposite action as the one articulated by the examiner (i.e., touch your toes if the examiner says to touch your head and vice versa).

HTKS involves six practice trials with feedback followed by ten test trials without feedback. Children who correctly execute five or more test trials then receive a second set of prompts: “touch your shoulders” and “touch your knees.” Once again, children have to perform the opposite action of the one orally indicated. This time, four practice trials with feedback are followed by ten test trials without feedback.

Each test trial is scored as 0 for an incorrect response, 1 for a corrected response (i.e., the child motions toward the body part congruent with the oral instruction but then quickly corrects to the incongruent body part), and 2 for a correct response. Children’s total scores reflected the sum of the six practice items plus the 20 test trials, which produce a possible range from 0 to 52 points. In the PRI dataset, the six practice trials were included to create a floor score for this relatively difficult task (Fuhs et al., 2013).

This task taxes all three elements of executive function: working memory to remember and act on the instructions, cognitive flexibility to switch among the body-tapping patterns, and inhibitory control to suppress the impulse to imitate the examiner’s movements. Four-year-old children’s performance on HTKS exhibits high ($r = .80$) test–retest reliability (Meador,

Turner, Lipsey, & Farran, 2013), and its convergent validity with teacher self-regulation reports ($r = .51$, $p < .01$) has been demonstrated (Ponitz et al., 2009).

The unstandardized means, standard deviations, skewness, and kurtosis values for the four assessment time points are listed below in table 13. The means increase over time, the standard deviations generally decrease, and the skewness and kurtosis values are mostly within acceptable ranges. The high skewness value at time four was treated through a latent variable transformation (see section 11.6).

Table 13: Heads-Toes-Knees-Shoulders descriptive statistics across time points

Item	Mean	SD	Skewness	Kurtosis
HTKS_t1	9.67	12.82	1.43	0.81
HTKS_t2	21.49	17.05	0.15	-1.46
HTKS_t3	36.61	13.59	-1.23	0.59
HTKS_t4	44.13	9.22	-2.47	7.23

Copy Design task

The Copy Design task (Osborne, Butler, & Morris, 1984) prompts children to recreate a geometric shape constructed by the examiner. If the child's shape meets the standards for imitative accuracy, then the child receives a score of 1 for that trial. If the child's shape does not resemble the examiner's shape, then the child receives a score of 0. The task involves eight geometric shapes that increase in difficulty across trials. For each shape, the child has two opportunities to replicate the design. Thus, the possible set of scores ranges between 0 and 16.

The Copy Design task assesses attentional control and persistence. As described by Cameron et al. (2012), the task requires children "to process visual information from an external stimulus, invoke a mental representation, and coordinate motor movements to reproduce the image" (p. 1240). A PRI report (Meador et al., 2013) found high ($r = .79$) test-retest reliability with the preschool children from the present sample. Moreover, confirmatory factor analytic studies indicate both its construct validity (Turner, Lipsey, Fuhs,

Vorhaus, & Meador, 2012) and predictive validity on subsequent academic achievement (Meador et al., 2013).

The unstandardized means, standard deviations, skewness, and kurtosis values for the four assessment time points are listed below in table 14. The means and dispersion values increase over time, and the skewness and kurtosis values are mostly within acceptable levels; the high skewness at time point one was treated through a latent variable transformation (see section 11.6).

Table 14: Copy Design descriptive statistics across time points

Item	Mean	SD	Skewness	Kurtosis
CD_t1	1.00	1.51	2.14	5.70
CD_t2	4.78	2.82	0.41	-0.11
CD_t3	7.78	2.83	-0.09	-0.26
CD_t4	9.06	3.00	-0.20	-0.29

Composite reliability and validity of the objective indicators

The internal consistency among the five standardized measures is .69 as assessed by Cronbach's alpha. Although just below Kline's (2011) standard of .7, other researchers (Berthoud & Geshuny, 2000) have argued that Cronbach's alpha scores above .6 are acceptable, especially for large sample studies using latent variables (Little, Nesselroade, & Linderberger, 1999). This is because latent variables account for the measurement error observed in the items, which are the source of the lower internal consistency values. By accounting for measurement error, the latent variables leave only the 'pure' variance over for analysis (see section 11.6 for more information about this study's latent variable approach).

Given this study's analytic sample size of 1140 respondents, and given this study's latent variable framework (see section 11.6), the Cronbach's alpha of .69 is considered acceptable. As for the validity of the combined indicators, the factor analysis section (section 11.6) describes the process that assesses whether the indicators are in fact measuring the self-regulation construct that they are intended to measure.

Summary of objective self-regulation measures

The task-based measures discussed above most directly capture the executive function component of self-regulation. That is, the indicators target children's cognitive control over their impulses, memory, and attention. By contrast, the informant-based self-regulation measures below target the "integration of these three executive functions into overt behavior" (Schmitt et al., 2015, p. 21). That is, the report-based measures discussed below target how children's executive function skills manifest as behavior in school.

11.5.2 Informant-report self-regulation measures

Informant-report measures scores derive from adult ratings of children's self-regulation. The first measure described below involves researcher reports of children's self-regulation, whereas the second involves teacher report.

Self-regulation assessor rating

The self-regulation assessor rating (SAR; Smith-Donald, Raver, Hayes, & Richardson, 2007) is a five-item, researcher-report scale ($\alpha = .95$) that assesses children's self-regulation within a classroom setting. PRI assessors completed the SAR for each child during the autumn and spring of pre-kindergarten as well as the spring of kindergarten and first grade. The measure's construct validity and convergent validity with several executive function indicators has been demonstrated (Smith-Donald et al., 2007).

The unstandardized means, standard deviations, skewness, and kurtosis values for the five SAR items are listed in table 15 below. The items are scaled from 0 – 3, where a score of 0 signifies low self-regulation and 3 signifies high self-regulation. The items show small but consistent increases in means versus slight decreases in dispersion across time; the skewness and kurtosis values are mostly within acceptable levels.

Table 15: Researcher-reported SAR descriptive statistics across time points

	Item	Mean	SD	Skewness	Kurtosis
Time 1	Item 1	2.42	0.80	-1.21	0.57
	Item 2	2.42	0.85	-1.45	1.32
	Item 3	2.35	0.82	-1.03	0.13
	Item 4	2.44	0.75	-1.23	0.85
	Item 5	2.27	0.88	-1.04	0.23
Time 2	Item 1	2.49	0.73	-1.34	1.13
	Item 2	2.49	0.74	-1.53	2.04
	Item 3	2.42	0.74	-1.13	0.68
	Item 4	2.55	0.69	-1.52	1.88
	Item 5	2.40	0.79	-1.18	0.69
Time 3	Item 1	2.69	0.61	-2.00	3.52
	Item 2	2.62	0.68	-1.84	2.94
	Item 3	2.59	0.67	-1.51	1.47
	Item 4	2.65	0.61	-1.67	2.15
	Item 5	2.66	0.62	-1.83	2.97
Time 4	Item 1	2.75	0.55	-2.27	4.98
	Item 2	2.67	0.61	-1.85	3.11
	Item 3	2.72	0.57	-2.05	3.71
	Item 4	2.78	0.48	-2.18	4.34
	Item 5	2.76	0.55	-2.61	7.14

Note: Item 1 = “Pays attention during instruction;” Item 2 = “Careful, interested in accuracy;” Item 3 = “Sustains concentration, willing to try repetitive tasks;” Item 4 = “Daydreams, has trouble focusing” (reverse-coded); Item 5 = “Distracted by sights and sounds” (reverse-coded)

Cooper-Farran behavioral rating scale

The Cooper-Farran behavioral rating scale (CBFRS; Cooper & Farran, 1988) is a 37-item scale ($\alpha = .92$) that assesses social skills and work-related skills in the classroom. In the present study, five items were selected from the work-related subscale given their conceptual overlap with the self-regulation construct (Fuhs et al., 2015). Those five items (see table 16) exhibit high internal consistency ($\alpha = .93$).

CFBRS items are on a 1 to 7 scale. In the original CFBRS scale, low scores indicate higher self-regulation. The items have been reverse-coded for this analysis to enhance interpretability (i.e., for the table 16 values below, higher item scores indicate higher levels of self-regulation). Each item has a set of behavioral descriptions for each odd-numbered item. For example, the minimum score on the “Independent work” item has the behavioral description “teacher prompting has only slight impact on work habits,” whereas the maximum score reflects that the child “works independently without supervision” (Cooper & Farran, 1988, p. 8).¹⁵

The unstandardized means, standard deviations, skewness, and kurtosis values of the CFBRS items are listed in table 16 below. The items show relatively consistent means and dispersion across time; the skewness and kurtosis values are all within acceptable levels as well.

¹⁵ Other commonly used subjective self-regulation measures include the Behavioral Rating Inventory of Executive Function (BRIEF), the Child Behavioral Questionnaire (CBQ), and the Child Behavioral Rating System (CBRS). One relevant difference between those surveys and the CFBRS is that the latter contains only classroom-related items. The classroom-specific nature of the CFBRS explains its use in this study as opposed to the other scales named in this footnote.

Table 16: Teacher-reported CFBRs descriptive statistics across time points

	Item	Mean	SD	Skewness	Kurtosis
Time 1	Item 1	3.43	1.82	0.32	-0.89
	Item 2	3.45	1.47	0.33	-0.04
	Item 3	3.32	1.46	0.30	-0.10
	Item 4	3.26	1.50	0.68	0.30
	Item 5	3.47	1.44	0.37	0.04
Time 2	Item 1	3.01	1.88	0.55	-0.87
	Item 2	2.97	1.55	0.54	-0.26
	Item 3	2.88	1.53	0.58	-0.22
	Item 4	2.89	1.50	0.72	0.15
	Item 5	3.01	1.55	0.55	-0.18
Time 3	Item 1	3.17	1.99	0.41	-1.10
	Item 2	3.04	1.72	0.60	-0.39
	Item 3	2.78	1.51	0.58	-0.21
	Item 4	2.75	1.45	0.83	0.59
	Item 5	3.01	1.72	0.61	-0.40
Time 4	Item 1	3.40	2.02	0.28	-1.18
	Item 2	3.11	1.75	0.49	-0.60
	Item 3	2.88	1.55	0.46	-0.49
	Item 4	2.75	1.47	0.76	0.36
	Item 5	3.19	1.79	0.47	-0.67

Note: Item 1 = “Behavior during designated work time;” Item 2 = “Compliance with teacher’s instructions relating to work;” Item 3 = “Completion of games and activities”; Item 4 = “Compliance with teacher’s instructions regarding behavior;” Item 5 = “Independent work”

11.5.3 Task-based versus informant-report measures: Which is best?

The PRI dataset includes both task-based and informant-report measures, so which one should be relied upon? Bronfenbrenner, whose bioecological model undergirds this

dissertation's analysis, asserts the importance of a multi-faceted view of human development (Bronfenbrenner & Morris, 2006). That is, although self-regulation may have objective indicators to measure it (e.g., children's DCCS scores), the child's and teacher's subjective experience of those self-regulation skills are also important:

Scientifically relevant features of an environment for human development not only include its objective properties but also the way in which the properties are subjectively experienced by the person living in that environment. [. . .] In the bioecological model, both objective and subjective elements are posited as driving the course of human development; neither alone is presumed sufficient. (Bronfenbrenner & Morris, 2006, p. 797)

The literature on self-regulation measures largely corroborates Bronfenbrenner's (2006) assertions regarding the dualistic approach. For example, McClelland (2012) found that teacher ratings of children's self-regulation skills can be subject to observer bias, whereas objective measures may lack applicability to classroom contexts because they are intended mostly for laboratory use (McClelland & Cameron, 2012). As Fuhs et al. (2015) explain, "direct assessments may capture children's available cognitive processes and teacher reports may assess how these processes are used in a real-world setting" (p. 217).

Overall, though, is one measure type more valid than the other? In a meta-analysis of both task-based and informant report self-regulation measures, Duckworth et al. (2012) found that, *ceteris paribus*, informant-report measures exhibit higher convergent validity with one another ($r = .54, p < .001$) than do task-based measures ($r = .15, p < .001$). However, the authors note that the task-based measures retain certain advantages; namely, they cannot be manipulated or distorted through personal bias (Duckworth & Kern, 2012). For example, if the informant-report assessors know that children have been assigned to the Tools condition, then the assessors may expect the children to have improved self-regulation skills, which could distort their scoring approach.

Ultimately, Duckworth and Kern (2012) conclude that "the optimal strategy is to use both task and questionnaire measures" (p. 268). The PRI data were chosen for this dissertation in part because they include a battery of task-based measures as well as both researcher- and teacher-reported self-regulation across four time points. In a previous PRI study (Fuhs et al., 2015), both the teacher-reported self-regulation ratings (CFBRS) and the task-based measures independently predicted gains in literacy and math. Thus, in line with several

studies (Duckworth & Kern, 2012; Fuhs et al., 2015; Gathercole et al., 2004; Tominey & McClelland, 2011), this dissertation includes both types of measures.

Although this study includes both task-based and informant-report measures, the two measure types were analyzed separately (i.e., as opposed to aggregating them into a single self-regulation construct). This is because the correlations between the two measure types are relatively modest (Blair, 2003; Duckworth & Kern, 2012), which suggests that they are capturing slightly different facets of the self-regulation construct (e.g., executive function versus the manifestation of executive function in classroom behavior). The construct validity of those task-based versus informant-report self-regulation measures was assessed using factor analysis, which is described below.

11.6 Factor analysis

Factor analysis assesses whether a large set of observed indicators actually reflects a smaller set of unobserved constructs called factors. For example, an intelligence test may contain 20 questions, but those questions may, overall, be targeting two distinct types of intelligence (e.g., mathematical intelligence and verbal intelligence). In order to identify those underlying constructs, factor analysis explores the observed inter-relationships among the indicators (i.e., their shared variance).

If the level of shared variance across the items is high, then this suggests that the items are actually capturing some shared underlying factor. The logic here is that if the scores on multiple items are usually in the same direction (i.e., they have a strong correlation), then those items are probably measuring the same construct.

By contrast, if two subsets of items exhibit strong inter-correlations within their subset scores but not between their subset scores, then this could be evidence that two distinct underlying factors are being captured by the full set of items. Overall, factor analysis contains two distinct approaches: Exploratory and confirmatory factor analysis, each of which is described in turn below.

11.6.1 Exploratory factor analysis

In exploratory factor analysis (EFA), the researcher makes no a priori assumptions about the factor structure underlying a set of indicators. That is, in the aforementioned intelligence test with 20 questions, the researcher would not necessarily expect those 20 items to load onto mathematical and verbal intelligence. Instead, EFA simply tests the inter-relationships among all items to identify how many underlying factors emerge.

In the present study, EFA was separately conducted for the five task-based executive function measures, the five teacher-reported self-regulations items, and the five researcher-reported self-regulation items. The EFA could reveal, for example, that the five task-based measures actually represent multiple distinct factors and therefore cannot be conceptualized as a singular latent executive function construct. The same logic applies to the teacher- and researcher-reported items. Thus, it is necessary to test the items' factor structure so that we confirm that we are measuring what we intend to measure.

In the present analysis, EFA models indicated that only one latent factor was underlying each set of items at each time point (see Appendix G for full output). That is, the five teacher-reported self-regulation items from the Cooper-Farran Behavioral Rating Scale (CFBRS; section 11.5.2) all loaded onto one latent factor; the five researcher-reported Self-Regulation Assessor Rating scale (SAR; section 11.5.2) all loaded onto one latent factor; and the five task-based self-regulation indicators (section 11.5.1) all loaded onto one latent factor. Thus, these analyses provided preliminary justification to assess self-regulation using three constructs: 1) a teacher-reported self-regulation construct, 2) a researcher-reported self-regulation construct, and 3) a task-based self-regulation, or executive function, construct.

Although the EFA provided preliminary evidence of self-regulation's latent factor structure, it was then important to employ confirmatory factor analysis (CFA) methods to corroborate that the analyses are capturing the self-regulation constructs that this study aims to analyze. The rationale and process for CFA models is presented in section 11.6.2 below.

11.6.2 Confirmatory factor analysis (CFA)

Similar to exploratory factor analysis (EFA), confirmatory factor analysis (CFA) assesses whether a set of indicators represents an underlying construct (e.g., intelligence, kindness, self-regulation) that is widely agreed to exist but is difficult to measure (Blunch, 2008). Despite that similarity, EFA and CFA are indeed different methods. With EFA, any group of indicators that exhibit interrelationships above a certain threshold (see Appendix G for details) classifies as a latent factor. That is, as its name suggests, EFA *explores* whether a factor exists in an ad hoc fashion whereby any indicators with sufficiently high inter-correlation emerge as a factor.

By contrast, CFA makes specific hypotheses about the underlying factor structure and tests those hypotheses against the data. That is, CFA, *ex ante*, identifies a set of indicators expected to form a construct, and then *confirms* whether that factor structure holds when examining the actual data.

In the present study, I employed CFA to assess whether my hypothesized factor structure aligned with the actual data. Specifically, the five task-based self-regulation measures were hypothesized to represent a single, underlying construct of executive function. The CFA could reveal whether all five indicators actually loaded onto the same construct, or, conversely, whether one or more items did not fit with the others.

Additionally, the items from the teacher-reported CFBR scale and the researcher-reported SAR scale were hypothesized to represent children's teacher- and researcher-reported self-regulation skills, respectively. Thus, CFA models test each factor's construct validity (Bryman, 2012): that is, does the set of indicators capture the single, underlying construct that they are intended to capture?

In addition to assessing construct validity, the CFA process also results in two key benefits that will each be described below: 1) they account for measurement error in the original indicators, and 2) they can transform non-continuous and non-normal manifest indicators into continuous, normally distributed latent variables (Kline, 2015).

First benefit of CFA: Accounting for measurement error

One benefit of CFA is that it accounts for measurement error (Kline, 2015). Section 11.5.1 indicated that the internal consistency across the five self-regulation indicators is .69; this suggests that approximately 31% of the variance across the indicators could be attributable to measurement error (Bryman, 2012). In CFA, only the variance common to all the executive function battery items is extracted and then fused into a unitary construct. This resultant construct should, theoretically, contain only the 'pure' shared variance among the indicators without measurement error (Kline, 2015).

By contrast, in PRI's Tools study using these data (Farran & Wilson, 2014), the researchers averaged the five task-based self-regulation indicators into an equally-weighted, standardized composite self-regulation score. This approach is problematic for two reasons: 1) it neglects measurement error, and 2) it falsely suggests that each of the five measures captures the executive function construct equally well (i.e., because PRI equally weighted the five scores to create the composite).

By contrast, CFA estimates a specific factor loading for each indicator to determine how much variance in each indicator will be consolidated into the latent factor. For example, some tasks may better reflect the executive function construct than others, which will be represented by the different factor loadings and the weights given to those indicators in the eventual latent construct. Thus, this study goes beyond the existing PRI study with these data to provide a more accurate picture of the executive function construct using CFA.

Second benefit of CFA: Creating continuous, normally-distributed variables

A second benefit of CFA is that it can transform non-normal, non-continuous observed indicators into normally distributed latent variables. As described in section 11.5.1, multiple task-based self-regulation measures exhibited problematic levels of skewness and kurtosis at various time points; that is, the scores were non-normally distributed. Through CFA, the non-normally distributed observed scores are transformed into normally distributed latent values (Kline, 2015, p. 12).

Given the conceptual issues surrounding the executive function construct (McClelland & Cameron, 2012), its associated measurement error problems (Miyake et al. 2000), and the problem of non-normally distributed indicators, many studies have also used CFA to test the validity of singular, latent self-regulation constructs from a variety of observed indicators (Denham, Warren-Khot, Bassett, Wyatt, & Perna, 2012; Miyake et al., 2000; Wiebe, Sheffield, & Nelson, 2011; Willoughby et al., 2014). The present analyses follows in that same tradition.

11.6.3 CFA results for the task-based measures

Mplus 7.3 (Muthén & Muthén, 2012) was used to execute a CFA of the five task-based executive function indicators at each of the four data collection time points. In order to test the fit of the five indicators onto a latent construct, conventional model fit indices and thresholds were used.

Specifically, Comparative Fit Index (CFI) values above .90 and .95 were considered acceptable and excellent, respectively (Bryne, 2012; Kline, 2015); Root Mean Squared Error of Approximation (RMSEA) and Standardized Root Mean Residuals (SRMR) values under .08 (Hu & Bentler, 1999) were both deemed acceptable (see Appendix H for a description of each model fit index's meaning). Finally, χ^2 p-values are nearly always significant in large sample studies such as this one, which is not considered a cause for concern (Bryne, 2012).¹⁶

The CFA achieved acceptable fit at each time point as shown in table 17 below. This signifies that the executive function indicators loaded onto the latent factor sufficiently well at each time point, which suggests construct validity. Previous studies (Wiebe, Espy, & Charak, 2008; Wiebe et al., 2011) regarding the latent executive function factor structure also support a one-factor executive function model during the early childhood years.

¹⁶ Although the fit indices each have their own threshold values, these cut-off points are somewhat arbitrary (Tabachnick & Fidell, 2013). As such, these threshold values will not be regarded as firm cut-off points but rather useful guides to the researcher. If the thresholds are upheld too strictly, then the probability of falsely rejecting a good model will increase (Blunch, 2008).

Although other studies (Miyake & Friedman, 2012; Miyake et al., 2000) indicate that executive function skills divide into discrete sub-constructs (i.e., inhibitory control, cognitive flexibility, and working memory) during adolescence and adulthood, developmental psychology research (Denham et al., 2012; A. Diamond, 2006) suggests that those sub-constructs are too inchoate to be differentiable in the early childhood years. Thus, the CFA results shown in table 17 below (and the EFA results shown in Appendix G) corroborated the one-factor approach, which will be employed in the upcoming analyses to be described in section 11.11.

Table 17: CFA results for the latent executive function construct across time points

	χ^2 (df)	CFI	SRMR	RMSEA
Time point 1	18.037 (5)*	.963	.032	.064
Time point 2	25.971 (5)*	.951	.034	.080
Time point 3	9.611 (5)	.988	.020	.037
Time point 4	9.791 (5)	.990	.018	.038

Note for this table and all subsequent tables: * = $p < .05$, ** = $p < .01$, *** = $p < .001$, and + signifies marginal significance of $p < .10$

11.6.4 CFA results for the informant-report self-regulation indicators

As described in section 11.5.3, the SAR and CFBRS will be modeled separately as researcher-reported and teacher-reported self-regulation constructs, respectively. Thus, table 18 indicates the fit indices for the latent researcher-reported SAR construct, whereas table 19 shows the fit indices for the latent teacher-reported CFBRS construct.

The values in both tables exhibit model fit indices that are within the model fit thresholds (i.e., CFI above .90; RMSEA and SRMR under .08), which, again, indicates construct validity. Consequently, all three self-regulation constructs are now continuous, normally distributed, and have been adapted to account for measurement error in preparation for the analyses.

Table 18: CFA results for the latent researcher-reported SAR construct across time points

	χ^2 (df)	CFI	SRMR	RMSEA
Time point 1	30.397 (5)*	.990	.011	.067
Time point 2	24.531 (5)*	.991	.012	.059
Time point 3	17.923 (5)*	.992	.010	.047
Time point 4	5.576 (5)	.999	.008	.009

Table 19: CFA results for the latent teacher-reported CFBR construct across time points

	χ^2 (df)	CFI	SRMR	RMSEA
Time point 1	10.363 (5)	.998	.008	.031
Time point 2	11.674 (5)*	.997	.008	.034
Time point 3	19.543 (5)*	.994	.009	.050
Time point 4	33.283 (5)*	.990	.013	.070

11.6.5 Measurement invariance

Because the self-regulation constructs were measured at four time points, it is necessary to assess the latent constructs' longitudinal consistency of measurement. That is, the CFA models indicate adequate fit for the three self-regulation constructs at each time point, but can we be sure that the indicators are measuring each construct the same way across time? Or might different indicators measure self-regulation differently across time points? If children's self-regulation were observed to increase over time, then it would be crucial to confirm that the increase resulted from genuine developmental growth rather than from measurement inconsistency in the self-regulation construct.

This can be tested using longitudinal measurement invariance models, which assess whether a set of indicators maintains the same factor structure over time (Kline, 2015). Measurement invariance assessment involves three progressively stricter tests: configural (i.e., same pattern of factor loadings across time points), weak (i.e., equal factor loadings across time points), and strong invariance (i.e., equal factor loadings and measurement intercepts across time points; see Appendix I for a more thorough description of the three invariance tests).

In line with the recommendations of Wu et al. (2010), the executive function indicator scores were standardized for the measurement invariance models. This is done because the metrics of children's task-based scores differed between measures (e.g., DCCS is on a 0 – 4 scale, whereas HTKS is 0 – 52); these issues make the measurement look inconstant over time even though it is not. As Wu explains, “The large difference in the range of subscales makes the cross-time comparison of the total scores impossible. It is not surprising if researchers find a poor model fit when examining strong invariance based on these observed scores with large range differences [. . .] the use of the variance and covariance structure of the Z scores makes the investigation of strong invariance sensible” (Wu, Liu, Gadermann, & Zumbo, 2010, p. 132).

By contrast, standardization is unnecessary for the subjective self-regulation measures because the constituent indicators for each are on the same scale with relatively small score ranges for both SAR (0 – 3) and CFBRs (1 – 7). Thus, the tables below illustrate the measurement invariance model fit indices for the executive function construct (table 20), SAR construct (table 21), and CFBRs construct (table 22).

The configural, weak, and strong measurement invariance model output in tables 20 through 22 show each self-regulation construct's acceptable measurement invariance values across the four time points. The χ^2 p-values for all measurement invariance models were significant, but, again, this is not considered a cause for concern given the other acceptable fit indices and the large sample size (Bryne, 2012).

Table 20: Configural, weak, and strong measurement invariance results for the executive function construct

	χ^2 (df)	CFI	SRMR	RMSEA
Configural	367.241 (134) ^{***}	.927	.073	.052
Weak	405.753 (146) ^{***}	.919	.076	.052
Strong	413.479 (161) ^{***}	.921	.076	.049

Table 21: Configural, weak, and strong measurement invariance results for the SAR construct

	χ^2 (df)	CFI	SRMR	RMSEA
Configural	293.029 (134) ^{***}	.987	.027	.032
Weak	336.790 (146) ^{***}	.985	.046	.034
Strong	679.556 (161) ^{***}	.959	.088	.053

Table 22: Configural, weak, and strong measurement invariance results for the CFBRs construct

	χ^2 (df)	CFI	SRMR	RMSEA
Configural	291.059 (134) ^{***}	.990	.021	.032
Weak	366.759(146) ^{***}	.986	.040	.036
Strong	706.729 (161) ^{***}	.966	.066	.055

11.6.6 Self-regulation measures summary

The PRI data were chosen for the present study given their diversity and quality of self-regulation measures. In order to further improve the construct validity of the raw measures, both the task-based and informant-report self-regulation measures were transformed into continuous and normally distributed latent variables that account for measurement error through CFA.

Subsequently, measurement invariance models demonstrated the longitudinal consistency in measurement of each latent construct across time. Thus, the self-regulation measures, which are the outcomes of interest, are now ready for analysis in the following chapter. The sections below outline (much more briefly) the two remaining measures for this study.

11.7 Tools activity implementation data

Study two examines which Tools activities are associated with children's self-regulation. The PRI dataset includes both teacher-reported implementation frequency data as well as researcher-observed implementation fidelity data for specific Tools activities. Each of those variables was used in the analysis, and each is described sequentially below.

11.7.1 Tools implementation frequency data

During the Tools implementation year (i.e., pre-kindergarten), teachers were surveyed during the autumn and spring regarding their frequency of Tools activity implementation. Teachers responded on a 0 – 3 scale, where 0 signifies that the teacher reported never implementing a specific activity, 1 signifies rarely, 2 signifies frequently, and 3 signifies daily implementation.

Although Tools includes 61 instructional activities in total, not all 61 activities are represented in the PRI dataset. Instead, the teacher-reported implementation measure includes 45 items. This is because some of the 45 items collapse a set of conceptually similar activities into a single item. For example, the Tools manual (Leong & Bodrova, 2011) enumerates ten distinct “Story Lab” literacy activities; however, in the teacher-report implementation measure, all ten are condensed into a single item that asks “How often did you do Story Lab Activities?”

Thus, most of the activities (74%) are individually represented in the teacher-reported implementation measure, whereas some are collapsed into categories (e.g., Story Lab, Mystery Activities). The internal consistency ($\alpha = .94$) of the teacher reports at each time point as well as the correlation ($r = .67, p < .01$) between time points were both high.

In line with Phillips et al. (2012), time one and time two reports were averaged to provide a more accurate picture of activity implementation across the year. If time one reports had been used on their own, then they would likely not capture accurate implementation levels because those reports derived from only the beginning of the school year, when teachers were just starting to use Tools with a new group of students.

By averaging the variables, a more precise picture of Tools implementation over the school year is obtained. Thus, the composite variable maintains the original 0 – 3 scales and includes both autumn and spring data. The descriptive statistics for the teacher-reported activity implementation data can be found in Appendix J.

11.7.2 Tools implementation fidelity data

Teachers' fidelity of Tools implementation was measured using an instrument created for the PRI study (Vorhaus & Meador, 2010). As described in section 11.2, the researchers collected fidelity data at three time points (autumn, winter, and spring of the pre-kindergarten year). To test fidelity, PRI used the Tools curriculum manual (Leong & Bodrova, 2011) to identify appropriate teacher actions for each activity.

For example, the Tools manual enumerates six steps for the buddy reading activity, which include steps such as “teacher releases a few children at a time” and “teacher scaffolds children’s oral language” (p. 343). Each Tools activity has a similar set of appropriate teacher steps associated with it, and the PRI fidelity measure quantifies the number of appropriate steps observed in teachers’ practice.

Given that each activity has a different number of steps, the possible fidelity score across activities varies widely (Min = 1, Max = 26). Thus, in order to compare fidelity scores, activity counts were first standardized to have a mean of zero and standard deviation of 1. This ensured that activities with more steps did not automatically receive higher fidelity scores (see Stechuk, 2008 for another Tools study that standardized its fidelity variables for the same purpose). Once the scores were standardized, the scores from the three time points were averaged into a composite variable (see Phillips et al., 2012).

Out of the 61 Tools activities, the PRI dataset had observations for 40 (versus 45 for the teacher-reported implementation frequency variable described in the previous section). The internal consistency of these composite variables was acceptable ($\alpha = .84$). The descriptive statistics for all researcher-reported Tools activity implementation data can be found in Appendix J.

11.8 Covariates

In order to isolate the unique effect of Tools activities on children's self-regulation, the analyses also controlled for the host of covariates described below.

Child age

Previous research indicates that self-regulation capacity is positively associated with age (Kopp, 1982; Moilanen, Shaw, Dishion, Gardner, & Wilson, 2009), whereby older children, on average, have better self-regulation capacity than younger children. Thus, all analyses controlled for child age in months at the first child assessment session (e.g., 53.5 months).¹⁷

Child gender

Previous research indicates that, *ceteris paribus*, females have more robust self-regulation skills than do males (Barnett et al., 2008; Kochanska et al., 2001). Thus, child gender (girls = 0, boys = 1) was included as a covariate in all analyses to determine whether Tools differentially impacted male and female students.

Child special educational needs status

Children with special educational needs (SEN) have been shown to have diminished self-regulation relative to their peers without special needs (Gulchak, 2008; Soares et al., 2009). In the United States, SEN students receive individualized education plans (IEP), which specify additional accommodations required by the child to meet his or her education needs. For example, an IEP may say that a child requires 50% more time in a testing environment than other students in the classroom. Thus, all analyses included children's SEN status (0 = child has no IEP, 1 = child has an IEP) as a covariate, which enables evaluation of Tools' differential effectiveness for SEN versus non-SEN students.

¹⁷ The analyses were also attempted with age as a time-varying covariate as well as a model with random effects for age. Those more complex approaches yielded similar results, so the most parsimonious model of fixed effects for age was ultimately selected (see Appendix N for my four tested approaches to treating the time variable).

English language learner status

English language learners (ELL) have been shown to encounter more academic and self-regulatory difficulties in educational contexts (Chularut & DeBacker, 2004). It is not the case that students who speak languages other than English have weaker self-regulation; rather, accessing education in one's non-native language may challenge the child's ability to academically and behaviorally excel (Hammer et al., 2012). Thus, all analyses included children's English language learner status (0 = not ELL, 1 = ELL) as a covariate to test Tools' differential effectiveness for ELL and non-ELL students.

Child socio-economic background

Children's socio-economic background is significantly associated with self-regulation development (Blair & Raver, 2014; Flouri et al., 2014; Raver et al., 2013). Unfortunately, the PRI data lack accurate socio-economic data because of the United States' Family Educational Rights and Privacy Act (FERPA) precluded such data collection. Thus, socio-economic background was not included as a covariate in this study or the previous PRI studies using these data (Fuhs et al., 2013; S. Wilson & Farran, 2012).

Nonetheless, according to a PRI study using the same data, "all children in this study came from public pre-k programs targeted to low-income families. Therefore, it can be assumed that most, if not all, children in the study were from low-income backgrounds" (Fuhs et al., 2015, p. 210). Thus, the results of this study bear on the question of whether Tools and its constituent activities effectively promote low-income children's self-regulation, which limits generalizability to that population.

Child ethnicity

The analyses did not control for child ethnicity because the data sample lacked accurate ethnicity data for the students (Farran & Wilson, 2014). Although schools reported students' ethnicities, previous studies using these data have also omitted ethnicity from the analyses due to "concerns about reporting idiosyncrasies in the data obtained from some schools" (Farran & Wilson, 2014, p. 17).

Pre-test score

Each analysis controlled for a child's pre-test scores on the target self-regulation construct. For example, when analyzing a child's executive function score at the second time point (i.e., spring of pre-kindergarten), the child's autumn pre-test score was included in the latent growth model (see section 11.11). This same procedure was followed for both task-based and informant-report self-regulation scores.

It is necessary to control for pre-test scores so that the unique effect of the independent variable on the child's score growth can be ascertained. Without controlling for the pre-test score, the results may indicate that certain Tools activities are associated with high scores at the spring post-test; however, it could have been that the children had high self-regulation scores from the beginning of the school year, and that high spring scores do not actually reflect growth but rather stasis. By controlling for pre-test score, the amount of growth, stasis, or decline from a baseline point can be accurately estimated.

11.9 Missing data treatment

In a large and longitudinal dataset such as this one, missing data are inevitable. In the present study, the average percentage of missing data points on the outcomes was 6.5% (SD = 3.5%, Min = 1.1%, Max = 10.9%). Given that the amount of missing data increased substantially across time points (T1 average = 1.3%, T2 = 5.9%, T3 = 7.7%, T4 = 10.9%), missing data treatment was necessary in order to maintain power and minimize bias (Bryman, 2012).

Missing data treatment options include (but are not limited to) listwise and pairwise deletion, single imputation, and multiple imputation (Enders, 2010). Listwise and pairwise deletion both exclude missing data from the analysis; such exclusion reduces statistical power and may introduce bias by only including cases with fully available data (Field, 2013).

By contrast, single imputation (e.g., imputing the mean value for all missing data points) underestimates the dispersion in the original data, which also introduces bias (Enders, 2010). Consequently, those methods were considered unacceptable for this study; instead, Bayesian multiple imputation emerged as the strongest option as described below.

11.9.1 Bayesian multiple imputation for the present study

Multiple imputation is arguably the most sophisticated and reliable missing data treatment (Graham, 2009; Sterne et al., 2009). Multiple imputation uses a Bayesian expectation-maximization algorithm, which contains two steps. In the expectation step, the variance-covariance matrix is used to produce a set of regression equations that predict missing values using observed values. In the maximization step, the regression equations are refined to produce more accurate estimates vis-à-vis the variance-covariance matrix. The model converges when the iteratively produced regression equations generate values that are not significantly different from one another across iterations (Enders, 2010).

Overall, multiple imputation is considered (Byrne, 2012; Enders, 2010) to be the least biased missing data treatment, especially for large datasets. In fact, standard errors of multiply imputed estimates are 7 – 40% smaller than those of listwise deletion (Enders, 2010). The number of imputed datasets should range at least between three and five in order to ensure adequate estimation of the original missing data (Enders, 2010). Thus, this analysis involved the imputation of five datasets using the Bayesian estimator in Mplus version 7.3 (Muthén & Muthén, 2012).

11.9.2 The resultant analytic sample for the present study

Of the original 1145 children in the sample, the Mplus imputation model excluded five who were missing all data on all measures of interest. Mplus cannot reliably impute missing data using available data if there are no available data for a child in the first place. Thus, the present study has complete data for all covariates and measures on 1140 of the 1145 children from the original sample (i.e., 99.6% of children from the original sample are represented).

Whereas self-regulation growth trajectories were assessed for all 1140 children in the analytic sample, study two also involves analyses that relate specifically to Tools activities. Thus, for the parts of study two that focus exclusively on Tools activities, only data from Tools students ($n = 646$) were used. By contrast, the control group students did not experience any Tools activities, so they were not included in the Tools activity analyses. The divide between the full-sample analysis and the Tools-only analysis becomes clear with the discussion of study two's research questions, which are enumerated below.

11.10 Research questions and hypotheses for study two

Whereas study one, the meta-analysis, reviewed 'Tools' effect at the curricular level across existing Tools evaluation studies, study two involves more granular analyses of specific Tools activities' associations with self-regulation development. In addition to evaluating Tools activities, though, the PRI data enable analysis of children's self-regulation growth trajectories during early childhood. Thus, the four research questions below range from the analysis of children's self-regulation developmental trajectories to Tools' effect on those developmental trajectories.

11.10.1 Research question one: How does children's self-regulation develop, on average, between the pre-kindergarten and first grade years?

Previous studies (Landry et al., 2014; Piotrowski et al., 2013) indicate that self-regulation capacity increases rapidly beginning around age three. Thus, Hypothesis 1 is that self-regulation will, on average, increase between the pre-kindergarten and first grade years. Some children will likely exhibit different trajectories. Some may increase steadily over time while others may fluctuate. Thus, Hypothesis 2 is that there will be significant developmental variability across children.

The association between a child's initial self-regulation status (i.e., the child's self-regulation score at the beginning of pre-kindergarten) and growth rate (i.e., how the child's self-regulation scores develop through first grade) is also germane to the first research question. For example, do children with high levels of self-regulation at the beginning of pre-kindergarten grow more rapidly through first grade than those with initially low levels of self-regulation? Or do those with high initial levels develop more slowly relative to those with low initial levels?

Whereas Skinner's (1998) 'launch model' posits that children with a high initial status in some capacity will be launched toward faster growth over time, empirical studies have not supported this theoretical model for self-regulation. In fact, previous self-regulation studies (Baron, 2014; Moilanen et al., 2009) have mostly indicated a negative association between

initial status and growth rate, whereby children with initially low levels of self-regulation exhibit steeper growth trajectories to catch up with their peers.

Thus, Hypothesis 3 is that children with high initial self-regulation will exhibit shallower self-regulation growth relative to children with low initial self-regulation. Conversely, those with low initial self-regulation will increase more quickly relative to those with high initial self-regulation levels. All 1140 children in the analytic sample (i.e., both Tools and non-Tools children) were represented in this analysis.

11.10.2 Research question two: Does Tools differentially affect children's self-regulation developmental trajectories vis-à-vis comparison curricula?

Given that Tools is explicitly inspired (Bodrova & Leong, 2007) by Vygotsky's cultural-historical theory (1962), which aims to promote children's higher mental functions such as self-regulation, it is hypothesized that children in the Tools condition will exhibit improved self-regulation skills relative to their control group counterparts. Whereas the meta-analytic results from Chapter Nine indicated an overall null effect for Tools' students' self-regulation, none of those studies accounted for measurement error in the self-regulation constructs.

Thus, in this study, which is the first to use a latent variable approach with Tools self-regulation data, I expect to observe improved self-regulation skills for children randomly assigned to the Tools program vis-à-vis children randomly assigned to the business-as-usual classrooms (Hypothesis 4). Because this analysis compares Tools and non-Tools students, data from all 1140 children in the analytic sample were used.

11.10.3 Research question three: Which, if any, Tools activity groupings (e.g., literacy, make-believe play, attention-focusing activities) predict children's self-regulation trajectories?

Using the Tools curriculum manual (Leong & Bodrova, 2011) as a guide, this analysis investigates the associations between blocks of Tools activities and children's self-regulation. The Tools manual contains six blocks of Tools activities: Literacy, math, make-believe play, attention-focusing activities, introduction activities, and science (see Appendix J for the list of Tools activities in each block).

For example, the Tools manual (Leong & Bodrova, 2011) includes a make-believe play block that consists of four activities: play planning, play practice, play centers, and play cleanup. Thus, for the purposes of this analysis, those four activities were averaged into a composite play block variable to test the association between the play block (as outlined in the manual) and children's self-regulation.

Although the manual indicates that all activities should theoretically hone self-regulation, some activity groupings are more strongly emphasized than others. Specifically, Vygotsky's cultural-historical theory (1962) identifies make-believe play as the critical driver of children's self-regulation (Fernyhough, 2009); similarly, the make-believe play activity block and the attention-focusing activities are identified as primary drivers of self-regulation growth by the curricular developers (Bodrova & Leong, 2007). In addition to those two Tools activity blocks, math instruction has generally been shown to predict self-regulation improvement (Bull & Scerif, 2010; Fuhs et al., 2015).

Thus, even if the analyses from research question two were to reveal that Tools has no overall effect on self-regulation, Hypothesis 5 is that the make believe-play, attention-focusing, and math activity blocks will predict significantly improved self-regulation skills, whereas literacy, introduction activities, and science will not. Because this analysis is specifically focused on Tools activities, only data from Tools children ($n = 646$) were used.

11.10.4 Research question 4: Which, if any, of the 61 Tools activities predict children's self-regulation trajectories?

Given that no previous study has investigated the associations between individual Tools activities and children's self-regulation, the present analysis features no specific hypotheses regarding each activity. Thus, this analysis is exploratory as opposed to confirming specific hypotheses about specific instructional activities.

Instead, the main text body of the results section in Chapter Twelve will examine specific Tools activities from any activity block (see research question three) that exhibits significant associations with children's self-regulation. For example, if the make-believe play activity block exhibits a significant association with self-regulation, then each of the four individual

activities that collectively comprise the make-believe play activity block will be shared in the text body of the results section. In so doing, this analysis reveals which constituent activities explain the overall association between a Tools activity block and children's self-regulation.

If an activity block, such as literacy, does not exhibit a significant association with self-regulation, then each of the individual Tools literacy activity results will still be shared in Appendix K in order to assess the full set of Tools activities. Once again, because this analysis is specifically focused on Tools activities, only data from Tools children ($n = 646$) were used.

11.10.5 Summary of research questions and hypotheses

These four research questions can all be addressed using the data sample and measures described in this chapter. Research questions one and two pertain to students in both the Tools and comparison classrooms; thus, the full analytic sample ($n = 1140$ children) was used for those questions. By contrast, research questions three and four pertain only to Tools activities, which only Tools children experienced; thus, the Tools-only sample ($n = 646$ children) was used for questions three and four.

For research questions three and four, I should reiterate that the PRI data include two activity implementation variables: 1) teacher-reported implementation frequency of the Tools activities (section 11.7.1), and 2) the researcher-observed implementation fidelity of the Tools activities (section 11.7.2). For the main analysis, I used the implementation frequency data, whereas the researcher-observed fidelity data was used as a sensitivity analysis to check the robustness of the main analysis.

I used the fidelity data as a robustness check, as opposed to using it for the main analysis, because the PRI researchers only observed Tools teachers for three lessons throughout the entire school year (Farran & Wilson, 2014). Thus, the fidelity data captures a small snapshot of teachers' practice, whereas the teacher report data asks teachers to reflect on their practice throughout the entire school year.

Given those differences, the PRI researchers also recommend (Meador, 2015) using the implementation frequency data instead of the fidelity data for analysis. Once again, I opted

to use the fidelity data for sensitivity analysis in order to assess the robustness of the main analyses with the implementation frequency data.

The activity analyses with the Tools implementation frequency data are only relevant to research questions three and four, whereas one and two involve a more general investigation of children's self-regulation growth. In order to analyze all four questions of children's self-regulation change over time, sophisticated statistical techniques are necessary (Kline, 2015; Singer & Willett, 2003). The following section describes study two's statistical approach: multilevel structural equation modeling.

11.11 Statistical approach: Multilevel structural equation modeling

Analyses of the four research questions involved multilevel structural equation modeling, which fuses two techniques: multilevel modeling and structural equation modeling. Firstly, multilevel modeling serves to analyze data that are nested in some way (e.g., students nested within classrooms). In traditional approaches such as regression and analysis of variance (ANOVA), the models assume all observations to be independent of one another. However, this assumption is often untenable in the real world. Students' self-regulation scores in the same classroom are not independent because they are affected by one another, by their teacher, and by their school.

Given the inherent dependency in the data used here (i.e., students nested in classrooms), the assumption of independence of observations is violated. Consequently, the standard errors of the model coefficients are underestimated (Tabachnick & Fidell, 2013), which increases the probability of a Type I error (i.e., the appearance of a statistically significant result where no such significant result exists in the real world). This is the same issue described for the multilevel meta-analysis in section 8.4.6.

Because the PRI data has students nested within teachers nested within schools, it is necessary to employ multilevel modeling, which does not treat data points within the same cluster as independent. Rather, multilevel models estimate the shared variation among observations from the same cluster (e.g., classrooms or schools) and then use regression techniques to identify predictors of that variation.

Next, because self-regulation is a complex and unobservable psychological construct (McClelland & Cameron, 2012; Whiteside & Lynam, 2001), latent variable structural equation modeling (SEM) emerges as a valuable analytic tool. SEM enables the analysis of latent factors, which are underlying constructs that are believed to exist but are difficult to observe and measure without error (e.g., self-regulation).

SEM's ability to model latent factors introduces another benefit of SEM over multivariate parametric tests such as ANOVA; namely, one assumption of ANOVA, regression, and other general linear models is that "all predictors are perfectly reliable (no measurement error)" (Kline, 2011, p. 32). By contrast, latent variable models, which account for measurement error, are ideal for addressing the measurement error issues posed by many self-regulation tests (Duckworth & Kern, 2012; Wiebe et al., 2011; Willoughby et al., 2014).

As was described in the confirmatory factor analysis section, the latent factors are created out of observed measure scores (e.g., HTKS, DCCS). Section 11.6 detailed how the multiple self-regulation indicators were fused into a single latent self-regulation construct. Now that the CFA and measurement invariance models have confirmed the latent factor structure of self-regulation, SEM and multilevel modeling can be concurrently employed to analyze the latent self-regulation construct using data that involve students nested within classrooms nested within schools.

The resultant statistical framework is called multilevel structural equation modeling (MSEM), which I used for this study. The MSEM framework is itself a family of techniques that contains several sub-techniques within it (Muthén & Muthén, 2012). The main MSEM technique used in the present study is latent growth modeling, which is described below.

11.11.1 Latent growth models

Duncan et al. (2009) designate latent growth modeling as "the most important and influential statistical revolution to have recently occurred in the social and behavioral sciences" (p. 979). Latent growth models investigate the predictors of intra-individual and inter-individual change on a target construct over time.

In traditional techniques such as repeated-measures ANOVA, inter-individual variation is treated as error variance, also known as “nuisance parameters” (T. Duncan & Duncan, 2009, p. 981). However, rather than being error variance, the variation in people’s growth trajectories across time has substantive import. That is, in latent growth models, predictors (e.g., Tools activities) can be used to predict intra-individual variation (e.g., changes in children’s self-regulation growth) and inter-individual variation (e.g., differences between children’s self-regulation growth) across time.

In addition to modeling intra- and inter-individual variation, latent growth models can simultaneously designate a variable as being both independent and dependent (Kline, 2015). For example, gender could be used as a predictor of task-based executive function scores, which could then be used to predict teacher-reported self-regulation CFBRs scores. In this case, children’s executive function score acts as both a dependent variable (when gender is regressed on it) and then an independent variable (when it is used as a predictor of teacher-reported self-regulation scores). By contrast, neither fixed nor random effects ANOVA models are able to provide such flexibility (T. Duncan & Duncan, 2009).

11.11.2 Multilevel nesting structure for the PRI data

All latent growth models are, by definition, multilevel models (Singer & Willett, 2003). This is because latent growth models nest time points within people to investigate intra-individual change over time. Data observations taken from the same children across time (e.g., HTKS scores from the same child over four testing administrations) are statistically dependent. That is, a child’s score on a test during kindergarten is correlated with his or her score during first grade. This data dependency explains part of the rationale behind the latent growth model, which is also known as the multilevel model for change (Singer & Willett, 2003).

Thus, at the least, latent growth models are two-level multilevel models with time points nested within people (see Appendix L for the equations underpinning the latent growth model in the present study). However, additional nesting levels can be added according to the data structure. In the present study, the PRI dataset contains four time points nested within 1140 children nested within 80 classrooms nested within 59 schools nested within six school districts nested within two states. This nesting structure underscores the rationale for

applying the bioecological model to the data; in Bronfenbrenner’s model, the child is nested within a series of progressively more macro structures, and the chronosystem (i.e., time) is nested within all levels.

Despite the complex nesting structure of the PRI data, multilevel models only incorporate nested levels when there is substantial variability in the outcome variable at that level. For example, students in different school districts likely vary in their self-regulation abilities. However, if most of that inter-district variability can be explained at lower nested levels, such as the classroom and school levels, then the school districts are omitted from the statistical model. This is because there is no additional inter-district self-regulation variance to explain above what is already explained by the classroom and school.

In the present study, preliminary analyses were carried out to determine the amount of self-regulation variance across classrooms and schools. The intra-class correlation (ICC) is a statistic that quantifies the variance attributable across levels. For example, an ICC value of .06 at the classroom level would signify that 6% of the variability in children’s self-regulation can be explained by inter-classroom differences rather than differences among children within a classroom. Table 23 shows the ICC values across classrooms for the five objective self-regulation measures.

Table 23: Intra-class correlation (children in classrooms) values for each of the objective self-regulation measures

Item	Time 1	Time 2	Time 3	Time 4
CB	0.049	0.045	0.078	0.047
DCCS	0.073	0.067	0.084	0.039
CD	0.029	0.079	0.024	0.040
HTKS	0.082	0.065	0.042	0.023
PT	0.078	0.057	0.014	0.041

Table 23 indicates that all values are below .10; for example, the value of .084 for DCCS time three signifies that 8.4% of the variability in children’s DCCS scores is explained by differences across classrooms (rather than within children). The original PRI researchers also reported low ICC values across classrooms and even lower ICC values across schools

(Fuhs et al., 2013). Although there is no agreed upon threshold for ICC values (Tabachnick & Fidell, 2013), some researchers (H. Goldstein, 2011; Mehta & Neale, 2005; Steenbergen & Jones, 2002) still recommend using multilevel models despite small ICC values to provide more conservative standard error estimates for regression coefficients.

Thus, the present study incorporated time points nested within children nested within classrooms despite relatively low classroom-level ICC values. The school, district, and state levels were omitted from the model because of near-zero variability at those levels (after accounting for the classroom-level variability).¹⁸ Of course, omitting the school level does not signify that children's self-regulation skills did not vary across schools. Instead, because these analyses pertain to the effects of teachers' instructional practices on children's self-regulation skills, the classroom level variance was deemed more pertinent to analyze than the school level (as well as district or state level).

In sum, the present study nests time points within children within classrooms. A central aim of the study is to explain the self-regulation variability observed within children, across children, and across classrooms based on Tools implementation. In order to do so, the section below explains how latent growth models are specified with intercept and slope parameters to precisely estimate self-regulation growth trajectories.

11.11.3 Intercepts and slopes in latent growth models

All latent growth models estimate the *initial status*, or intercept, of a construct at a specific time point (e.g., self-regulation at the beginning of the school year), as well as the *growth rate*, or slope, across time points. The intercept can be set at different time points depending on the substantive research question at hand.

The intercept and slope parameters for the first research question were estimated using an unconditional growth model (UGM). The UGM explores how an outcome variable (e.g., self-regulation) changes over time without considering predictors (e.g., Tools activities) or

¹⁸ I attempted the four-level model with time points nested in children nested in classrooms nested in schools despite the low ICC values at the school level, but the model does not converge in Mplus. This is because Mplus cannot compute school-level estimates of explained variance when there is virtually no variance to explain (Muthén & Muthén, 2012).

covariates (e.g., child gender). It is called an ‘unconditional’ growth model because the self-regulation growth is not predicted, or conditioned, by any predictors or covariates. The predictors and covariates are excluded in order to observe how self-regulation develops, on average, among all children (Singer & Willett, 2003); thus, the UGM is ideally suited to address research question one regarding average self-regulation growth over time. The UGM is depicted below in figure 17.

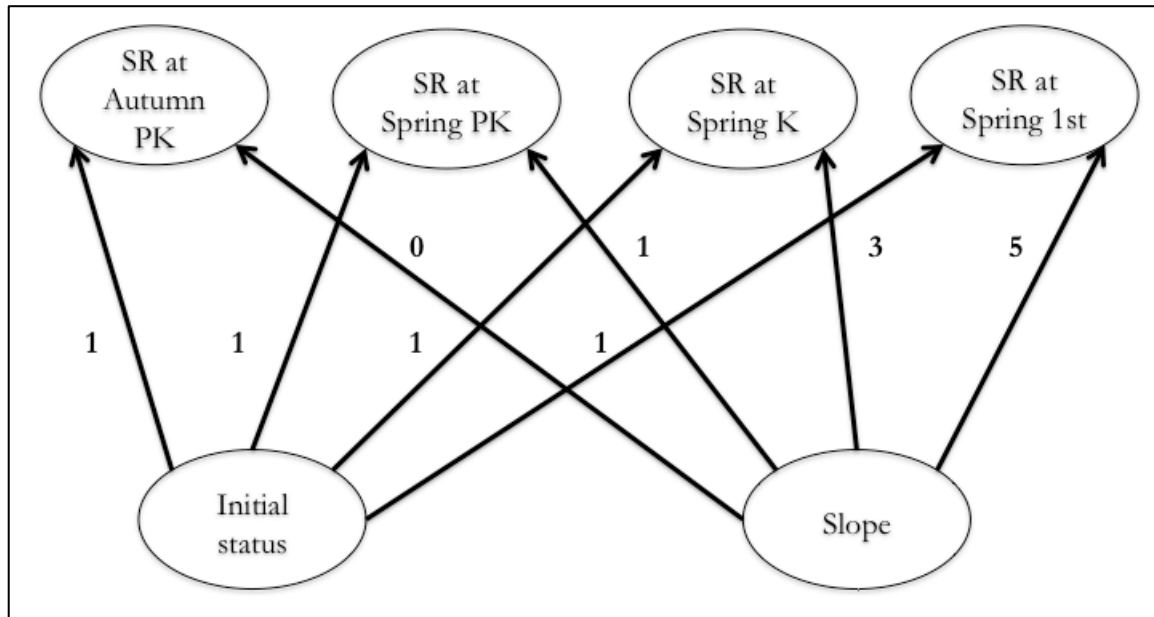


Figure 17: Unconditional growth model with a latent intercept, slope, and self-regulation at four time points

As figure 17 shows, the analysis for research question one designated the intercept at the first time point (i.e., pre-kindergarten autumn when children’s mean age = 54 months). This is because the question aims to investigate children’s self-regulation growth from the beginning of pre-kindergarten to the end of first grade, so pre-kindergarten autumn should be set as the intercept.

As for growth rates over time, the slope parameters were specified according to the discrete time points associated with the self-regulation testing administrations (see section 11.2 for a review of the project timeline). Specifically, the slope parameter was set with one school year between time point one and two (i.e., autumn to spring) versus one calendar year between time points two and three (i.e., pre-kindergarten spring to kindergarten spring).

This results in the following slope specifications: The first two growth parameters for time points one and two (i.e., beginning and end of pre-kindergarten) were set at 0 and 1, respectively, to represent a temporal distance of approximately six months of school. This is because the first self-regulation data were taken during October of pre-kindergarten, whereas the second time point data were taken approximately six months later in April and May.

The distance between times point two and three, however, was one calendar year (i.e., pre-kindergarten spring to kindergarten spring). Thus, the third growth parameter was set at 3, which is twice as large as the difference between 0 and 1. That doubling reflects the fact that the timing between testing administrations was twice as long (i.e., one calendar year versus six months of school). The fourth growth parameter was set at 5, which, again, is twice the distance between the 0 and 1 specifications at time points one and two.

Thus, the model in figure 17 captures children's average initial self-regulation status at the beginning of data collection as well as their average growth over the three-year study. This slope parameterization logic is in line with the recommendations of the Mplus software handbook (Muthén & Muthén, 2012).

By contrast, for research questions two through four, the analytic aim is to estimate Tools' effect on children's self-regulation. Because children experienced Tools throughout the pre-kindergarten year, the intercept should be placed at the second time point (i.e., the end of the pre-kindergarten year when the children's mean age = 60 months). In so doing, the substantive interpretation of the intercept is Tools' effect on children's self-regulation after a full school year of Tools implementation.

Conversely, if the intercept had been again placed at the first time point, then the intercept's value would be meaningless; it would estimate the effect of Tools before Tools had actually been implemented. Thus, the intercept was designated at the beginning of pre-kindergarten for research question one and the end of pre-kindergarten for questions two through four (see figure 18 for a graphic representation).

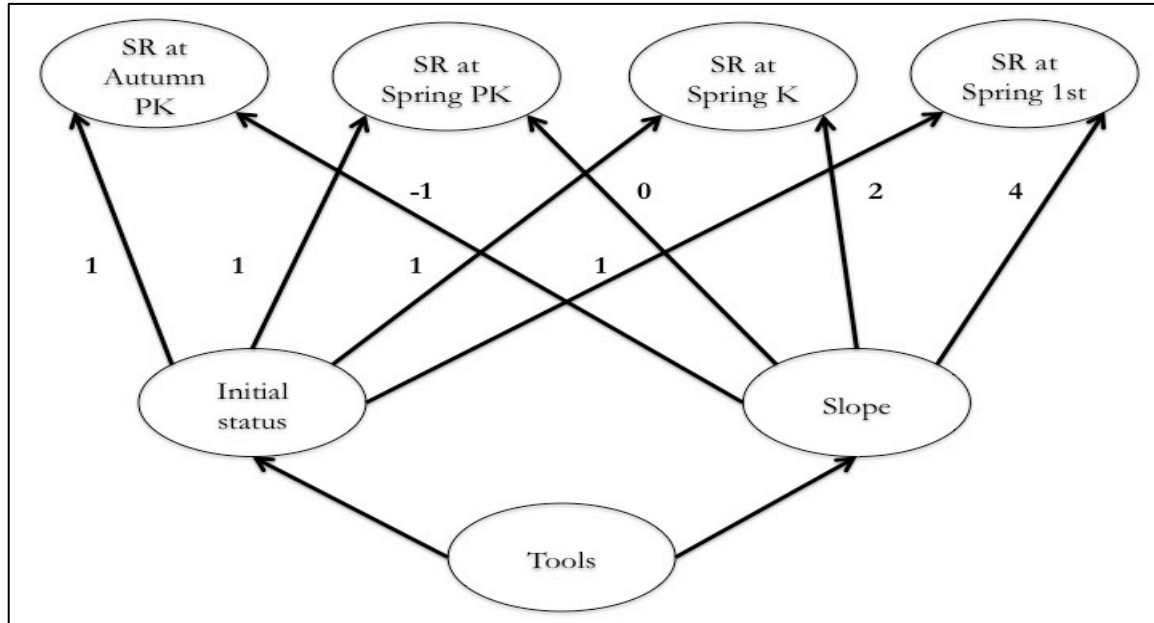


Figure 18: Conditional growth model for research questions two (child covariates such as gender are omitted to enhance graphical interpretability)

As the example in figure 18 shows, predictors (e.g., Tools) are added to the model to produce conditional growth models (i.e., the self-regulation change over time is conditioned, or predicted, by some covariate(s)). This process enables comparative analysis of model fit as the latent growth model becomes more complex (i.e., as more variables are added to it). Once the model is complete, it estimates self-regulation change across time as well as the predictors of that change. Specifically, in this study, the analysis targets whether Tools on the whole, as well as specific Tools activities, predict self-regulation change over time (see Appendix M for sample Mplus syntax and more technical explanation of the model).

11.12 Methodological summary

This chapter described study two's data sample, measures, ethics, and data analysis strategy. In sum, study two analyzes secondary data obtained through the Peabody Research Institute (PRI) at Vanderbilt University. The data derive from a large-scale randomized trial of Tools involving 1145 children in 80 classrooms in the United States. Using the rich self-regulation and Tools implementation data, study two investigates children's self-regulation growth trajectories as well as whether and how Tools affects that growth.

In order to conduct such analyses, I employed multilevel structural equation modeling techniques with a specific focus on latent growth modeling. The methodological complexity is warranted given the nature of the research aims – that is, predicting change among three latent self-regulation constructs among children nested within classrooms over four data collection points. The following chapter shares the results regarding self-regulation change and the predictors of that change between the pre-kindergarten and first grade years.

CHAPTER 12: Results for study two

This chapter presents the results for study two, which investigated Tools of the Mind's (Tools) overall effect on self-regulation as well as the associations between discrete Tools activities and self-regulation. Specifically, this chapter shares the results from the four research questions shared in the methodology chapter:

- How does children's self-regulation develop, on average, between pre-kindergarten and first grade?
- Does Tools differentially affect children's self-regulation developmental trajectories vis-à-vis comparison curricula?
- Which, if any, Tools activity groupings (e.g., literacy, make-believe play, attention-focusing activities) predict children's self-regulation trajectories?
- Which, if any, of the Tools activities predict children's self-regulation trajectories?

The results for each question are now shared sequentially below.

12.1 Research question one: Children's self-regulatory development between pre-kindergarten and first grade

Before assessing Tools' effect on children's self-regulation development, it is first important to understand how self-regulation develops, on average, within and between children. To address this question, unconditional growth models (UGM) are used to estimate children's average initial self-regulation status and growth trajectories without any predictors or covariates in the model.

Additionally, the UGM estimates the association between the initial status and slope; that is, do children with a high initial self-regulation status grow faster in self-regulation over time, or do those with initially low self-regulation grow faster to catch up with their peers? The initial status, or intercept, was set at the first time point (i.e., beginning of the pre-kindergarten year), and the slope parameters for the four time points were set at 0, 1, 3, and 5, respectively, to reflect the data collection timeline as was described in section 11.11.3.

The data sample for the present study includes seven self-regulation measures: the five task-based measures and the two informant-report measures. The UGM results for each are presented below. Importantly, the UGM results depict the growth trajectories for the Tools and control students combined ($n = 1140$) to analyze how self-regulation develops, on average, among children between the beginning of pre-kindergarten (approximately age four) and the end of first grade (approximately age seven).

After that, the research question two section (12.2) divides the sample by Tools versus comparison group children in order to assess developmental differences over time. First, however, the sections below present the UGM results for all children's growth on each of the seven self-regulation measures in turn.

12.1.1 Corsi Blocks

Model fit: Once again, the target model fit thresholds are RMSEA and SRMR values below .08, CFI values above .90, and, ideally, χ^2 p-values above .05. However, in large sample studies, the χ^2 p-value is nearly always below .05, which should not be a cause for concern (Bryne, 2012; T. Duncan, Duncan, & Strycker, 2006). Based on the target threshold indices, the Corsi blocks UGM exhibited excellent model fit ($\chi^2_{[2]} = 1.821$; $p = .39$, RMSEA = .008, CFI = .999, SRMR = .011).

Initial status and growth rate estimates: The unstandardized initial status and growth rate estimates are presented below in table 24.

Table 24: Unstandardized intercept and growth rate estimates for the Corsi Blocks task

	Estimate	S.E.	Estimate/S.E.	p-value
Intercept	1.139	0.034	33.947	0.001
Slope	.340	0.066	5.126	0.039
Quadratic	.073	0.021	3.560	0.050

Table 24 indicates that children's Corsi Blocks scores grew significantly over time. Specifically, the growth rate includes a significant ($b = .34$, $p < .05$) linear trend (i.e., the slope value) and also a positive quadratic trend ($b = .073$, $p = .05$), which signifies that children's

scores increase more rapidly across time points. The average Corsi Blocks growth trajectory between autumn of pre-kindergarten (time point one), spring of pre-kindergarten (time point two), spring of kindergarten (time point three), and spring of first grade (time point four) is depicted below in figure 19.

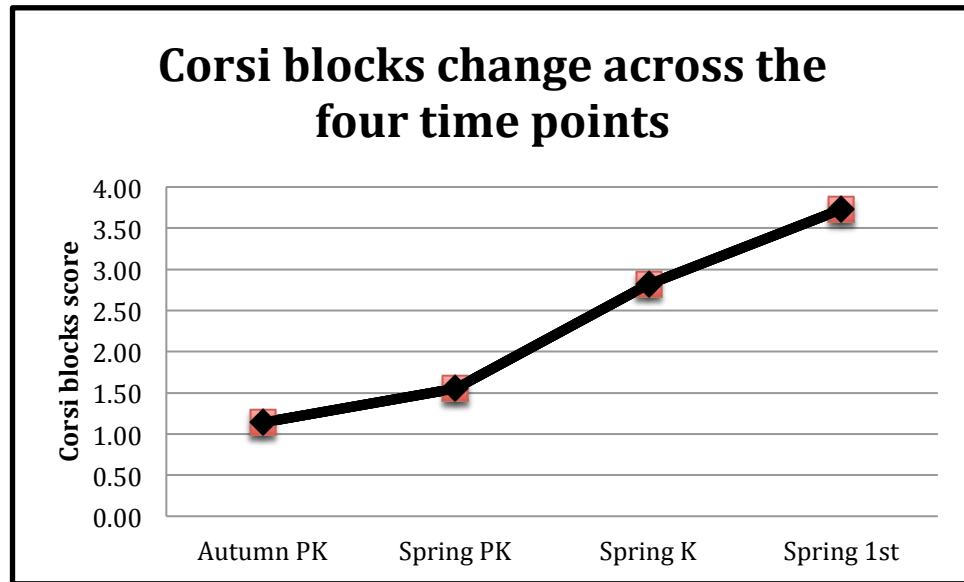


Figure 19: Average Corsi Blocks growth trajectory across time

Variation in observed results: The model results also indicated significant variability in both the intercept ($s^2 = .333$, $p < .001$) and linear slope ($s^2 = .105$, $p < .001$) across the 1140 children in the study sample. In line with the recommendations of Duncan and Duncan (2009), the quadratic variance was fixed to zero to improve model estimation. Nonetheless, the significant intercept and linear slope variance indicate that some children started high and stayed high, others started high and decreased over time, while others started low and stayed low, and so on. Thus, Hypothesis 2 regarding significant variability of developmental trajectories across children was verified.

Association between initial status and growth rate: The model indicated no significant association ($r = -.068$, $p = .606$) between the initial Corsi Block status and growth rate. This does not support Hypothesis 3 that children with higher initial self-regulation status would exhibit more shallow growth trajectories across time.

12.1.2 Heads-Toes-Knees-Shoulders

Model fit: The unconditional growth model exhibited excellent model fit ($\chi^2_{[1]} = 3.796$; $p = .054$, RMSEA = .051, CFI = .998, SRMR = .012).

Initial status and growth rate estimates: The unstandardized initial status and growth rate estimates are presented below in table 25.

Table 25: Unstandardized initial status and growth rate estimates for the HTKS task

	Estimate	S.E.	Estimate/S.E.	p-value
Intercept	9.704	.382	25.393	0.001
Slope	12.126	.307	39.447	0.001
Quadratic	-1.051	.056	-18.830	0.001

Table 25 indicates that children's HTKS scores grew significantly over time. Specifically, the growth rate includes a significant linear trend ($b = 12.126$, $p < .001$) but also a negative quadratic trend ($b = -1.051$, $p < .001$), which signifies that children's growth decelerates across time points. The average HTKS growth trajectory between the beginning of pre-kindergarten (time point one) and the end of first grade (time point four) is depicted below in figure 20.

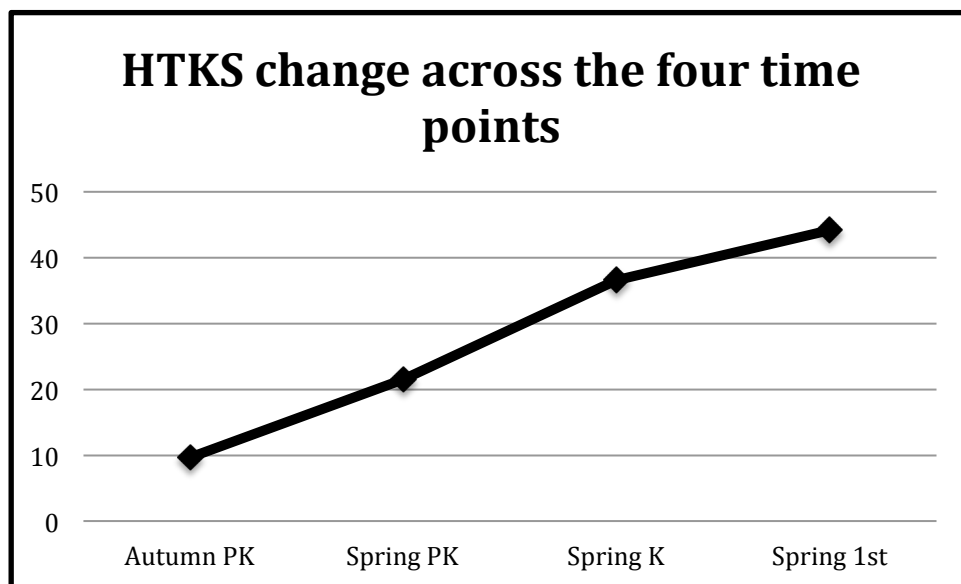


Figure 20: Average HTKS growth trajectory over time

Variation in observed results: The model results indicated significant variability in both the intercept ($s^2 = 150.123$, $p < .001$) and linear slope ($s^2 = 86.199$, $p < .001$) in the developmental trajectories across the 1140 children. Thus, Hypothesis 2 regarding significant variability of developmental trajectories across children was verified.

Association between initial status and growth rate: The model indicated a significant and negative association ($r = -.363$, $p < .001$) between initial HTKS status and growth rate. This supports Hypothesis 3 that children with higher initial status would exhibit more shallow growth trajectories across time.

12.1.3 Dimensional Card Sorting task

Model fit: The unconditional growth model exhibited excellent model fit ($\chi^2_{[2]} = 7.952$; $p < .05$, RMSEA = .050, CFI = .986, SRMR = .020).

Initial status and growth rate estimates: The unstandardized initial status and growth rate estimates are presented below in table 26.

Table 26: Unstandardized initial status and growth rate estimates for the DCCS task

	Estimate	S.E.	Estimate/S.E.	p-value
Intercept	1.287	.017	73.956	0.001
Slope	.422	.029	14.447	0.001
Quadratic	-.061	.009	-6.798	0.001

Table 26 indicates that children's DCCS scores grew significantly over time. Specifically, the growth rate includes a significant linear trend ($b = .422$, $p < .001$) but also a negative quadratic trend ($b = -.061$, $p < .001$), which signifies that children's growth decelerates across time points. The average DCCS growth trajectory between the beginning of pre-kindergarten (time point one) and the end of first grade (time point four) is depicted below in figure 21.

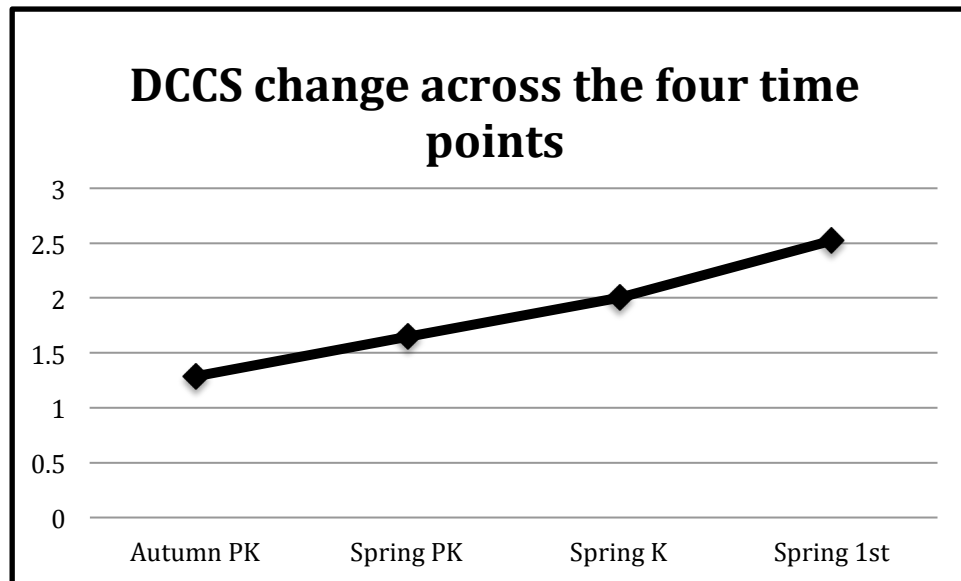


Figure 21: Average DCCS growth trajectory over time

Variation in observed results: The model results indicated significant variability in the initial DCCS status ($s^2 = .101$, $p < .001$) but not the growth rate ($s^2 = .002$, $p = .503$). Thus, Hypothesis 2 regarding significant variability of developmental trajectories across children was only partially verified. Children had different DCCS levels at the beginning of pre-kindergarten, but the pattern of their growth did not significantly vary across time.

Association between initial status and growth rate: The model indicated no significant association ($r = .082$, $p = .410$) between initial status and growth rate. This does not support Hypothesis 3 that children with higher initial status would exhibit more shallow growth trajectories across time.

12.1.4 Copy Design

Model fit: The unconditional growth model exhibited relatively poor model fit ($\chi^2_{[2]} = 47.704$; $p < .001$, RMSEA = .109, CFI = .952, SRMR = .049). Specifically, the CFI and SRMR are both within their threshold values (above .90 and below .05, respectively), whereas the RMSEA is not (.109 versus a threshold of .08).

The explanation behind the RMSEA's unacceptable value may be two-fold: 1) the measure is inherently positively biased, meaning that its value tends to be higher than the actual model fit, and 2) the size of that positive bias is linked to small degrees of freedom values

(Kline, 2015). Given that this model has only two degrees of freedom, the amount of positive bias may be unsurprising. Nonetheless, the unconditional growth model results for the Copy Design task should be viewed with additional caution.

Initial status and growth rate estimates: The unstandardized initial status and growth rate estimates are presented below in table 27.

Table 27: Unstandardized initial status and growth rate estimates for the Copy Design task

	Estimate	S.E.	Estimate/S.E.	p-value
Intercept	.994	.045	22.191	0.001
Slope	4.517	.110	41.435	0.001
Quadratic	-.777	.036	-21.377	0.001

Table 27 indicates that children's Copy Design scores grew significantly over time. Specifically, the growth rate includes a significant linear trend ($b = 4.517$, $p < .001$) but also a negative quadratic trend ($b = -.777$, $p < .001$), which signifies that children's growth decelerates across time points. The average Copy Design growth trajectory between the beginning of pre-kindergarten (time point one) and the end of first grade (time point four) is depicted below in figure 22.

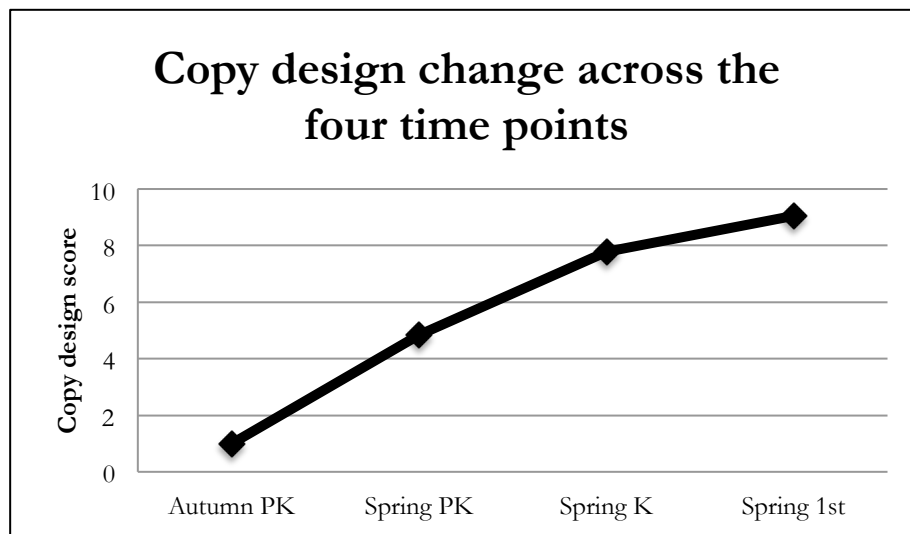


Figure 22: Average Copy Design growth trajectory across time

Variation in observed results: The model results indicated significant variability in both the intercept ($s^2 = 1.872$, $p < .001$) and linear slope ($s^2 = .517$, $p < .001$) across the 1140 children in the sample. Thus, Hypothesis 2 regarding variability of developmental trajectories across children was verified.

Association between initial status and growth rate: The model indicated no significant association ($r = -.071$, $p = .407$) between initial status and growth rate. This does not support Hypothesis 3 that children with higher initial status would exhibit more shallow growth trajectories across time.

12.1.5 Peg Tapping

Model fit: The unconditional growth model exhibited relatively poor model fit ($\chi^2_{[1]} = 43.808$; $p < .001$, RMSEA = .194, CFI = .945, SRMR = .042). Specifically, the CFI and SRMR are both within their threshold values (above .90 and below .05, respectively), whereas the RMSEA is not (.19 versus a threshold of .08). As with the Copy Design task, the explanation behind the high value may result from the fact that RMSEA values are positively biased with few degrees of freedom (Kline, 2015). Nonetheless, the unconditional growth model results for the Peg Tapping task should be viewed with additional caution.

Initial status and growth rate estimates: The unstandardized initial status and growth rate estimates are presented below in table 28.

Table 28: Unstandardized initial status and growth rate estimates for the Peg Tapping task

	Estimate	S.E.	Estimate/S.E.	p-value
Intercept	4.763	.170	28.065	0.001
Slope	4.220	.103	40.801	0.001
Quadratic	-.447	.017	-25.857	0.001

Table 28 indicates that children's Peg Tapping scores grow significantly over time. Specifically, the growth rate includes a significant linear trend ($b = 4.22$, $p < .001$) but also a negative quadratic trend ($b = -.447$, $p < .001$), which signifies that children's growth decelerates across time points. The average Peg Tapping growth trajectory between the

beginning of pre-kindergarten (time point one) and the end of first grade (time point four) is depicted below in figure 23.

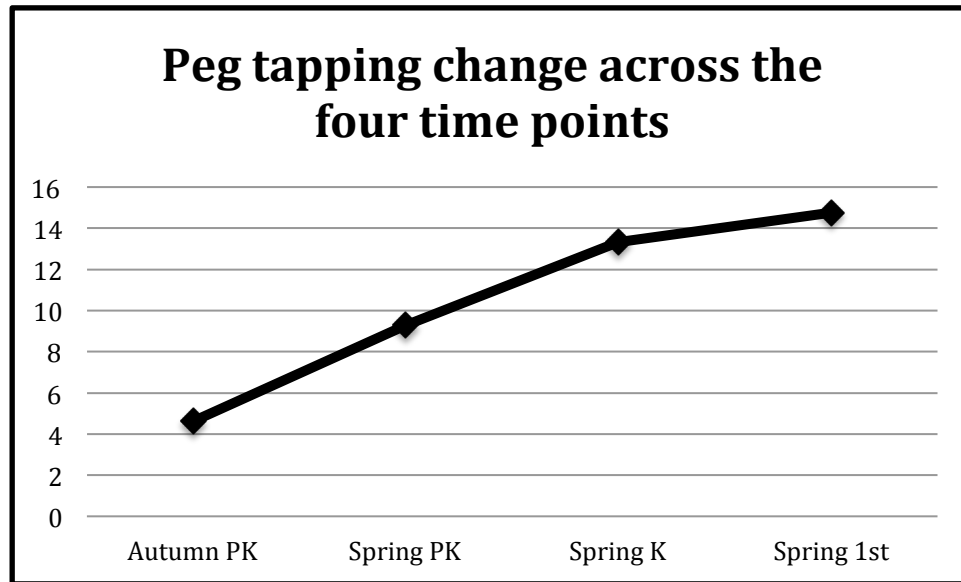


Figure 23: Average peg tapping growth trajectory across time

Variation in observed results: The model results indicated significant variability in both the intercept ($s^2 = 23.329$, $p < .001$) and linear slope ($s^2 = 7.976$, $p < .001$) across the 1140 children in the sample. Thus, Hypothesis 2 regarding variability of developmental trajectories across children was verified.

Association between initial status and growth rate: The model indicated a significant and negative association ($r = -.569$, $p < .001$) between the initial Peg Tapping status and growth rate. This supports Hypothesis 3 that children with higher initial status would exhibit more shallow growth trajectories across time.

12.1.6 Researcher-reported self-regulation (SAR)

Model fit: The unconditional growth model exhibited excellent model fit ($\chi^2_{[161]} = 500.567$; $p < .001$, RMSEA = .043, CFI = .973, SRMR = .059).

Initial status and growth rate estimates: In this model, the five SAR indicators have already been fused into a singular latent construct, which is scaled to have a mean of zero

(Little, Slegers, & Card, 2006). Thus, the intercept and growth rates are not in the original 0 – 3 scale of the SAR indicators. The intercept of the latent variable is estimated to be zero, and its standard error and significance level are not estimated. Thus, table 29 below simply shows the unstandardized growth rate of the latent SAR construct.

Table 29: Unstandardized growth rate for the SAR construct

	Estimate	S.E.	Estimate/S.E.	p-value
Slope	.704	0.053	13.330	0.001
Quadratic	-.004	0.017	-10.441	0.249

Table 29 indicates that children’s SAR scores grow significantly over time. Specifically, the growth rate includes a significant linear trend ($b = .704$, $p < .001$). The quadratic growth parameter was not significant ($b = -.004$, $p = .249$) and was thus omitted from subsequent models involving SAR. The average SAR growth trajectory between the beginning of pre-kindergarten (time point one) and the end of first grade (time point four) is depicted below in figure 24.

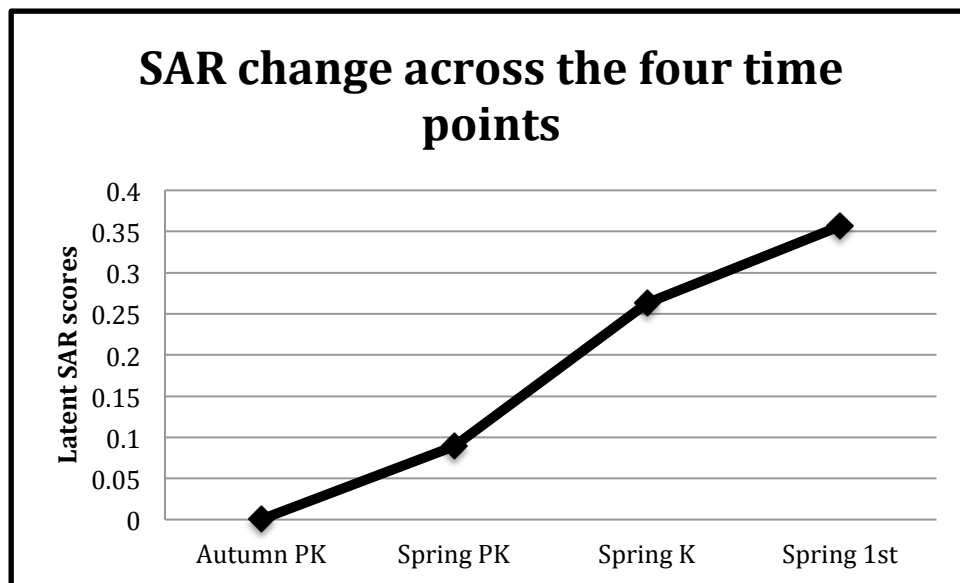


Figure 24: Average SAR growth trajectory across time

Variation in observed results: The model results indicated significant variability in both the initial status ($s^2 = .216$, $p < .001$) and growth rates ($s^2 = .011$, $p < .05$) across the 1140

children in the sample. Thus, Hypothesis 2 regarding variability of developmental trajectories across children was verified, although the variability in initial status was more notable than in the growth rates.

Association between initial status and growth rate: The model indicated a significant and negative association ($r = -.586$, $p < .001$) between initial SAR status and growth rate. This supports Hypothesis 3 that children with higher initial status would exhibit more shallow growth trajectories across time.

12.1.7 Teacher-reported self-regulation (CFBRS)

Model fit: The unconditional growth model exhibited excellent model fit ($\chi^2_{[160]} = 522.907$; $p < .001$, RMSEA = .045, CFI = .977, SRMR = .051).

Initial status and growth rate estimates: As in the SAR section, the five CFBRS indicators have already been fused into a singular latent construct with a mean of zero. Thus, the initial status and growth are not in the original 1-7 metric of the CFBRS indicators. Table 30 below simply shows the unstandardized growth rates of the CFBRS construct.

Table 30: Unstandardized growth rate estimates for the CFBRS construct

	Estimate	S.E.	Estimate/S.E.	p-value
Slope	.704	0.022	-3.339	0.017
Quadratic	-.178	0.005	3.008	0.031

Table 30 indicates that children's CFBRS scores include a significant linear trend ($b = .704$, $p < .05$) and a negative quadratic trend ($b = -.178$, $p < .05$), which signifies that children's scores decelerate across time points. The average CFBRS growth trajectory between the beginning of pre-kindergarten (time point one) and the end of first grade (time point four) is depicted in figure 25.

Figure 25 indicates that teacher-reported self-regulation skills increase during pre-kindergarten, but the teacher ratings actually begin to decrease between the spring kindergarten and spring first grade time points. This may be because teachers have

heightened expectations for students' self-regulatory behaviors as the children age, but this hypothesis could not be tested with the available data.

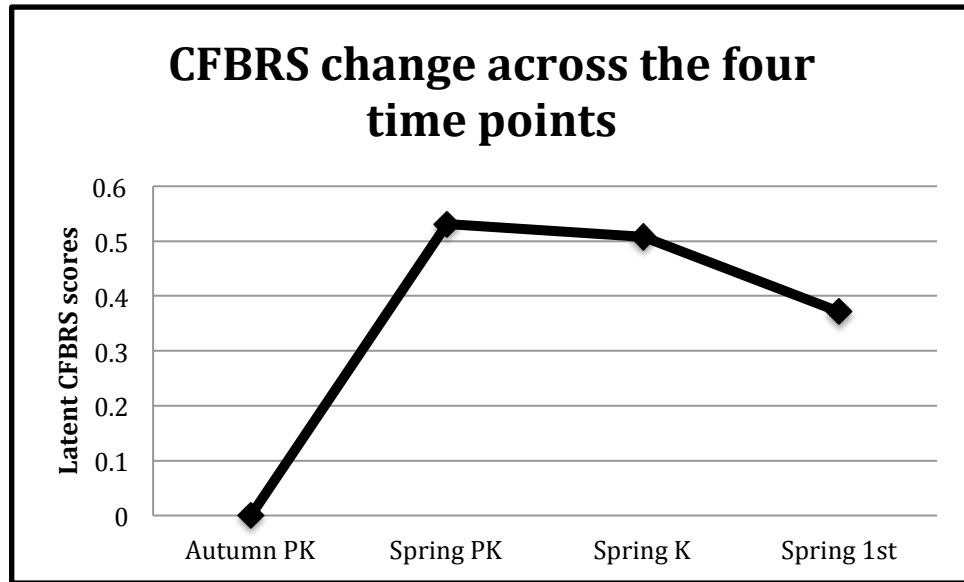


Figure 25: Average CFBRS growth trajectory across time

Variation in observed results: The model results indicated significant variability in both the intercept ($s^2 = 2.023$, $p < .001$) and linear slope ($s^2 = .159$, $p < .001$) across the 1140 children in the sample. Thus, Hypothesis 2 regarding variability of developmental trajectories across children was verified.

Association between initial status and growth rate: The model indicated a significant and negative association ($r = -.397$, $p < .001$) between initial CFBRS status and growth rate. This supports Hypothesis 3 that children with higher initial status would exhibit shallower, or, as observed, negative growth trajectories across time.

12.1.8 Summary on self-regulation development

The 1140 children in the analytic sample exhibited, on average, significant and positive growth between pre-kindergarten and first grade on the self-regulation indicators. In the Copy Design and SAR measures, growth appeared to be mostly monotonic; that is, a relatively linear average growth trajectory was observed whereas a significant quadratic trend

was not. The five remaining self-regulation measures all had significant quadratic trends, which suggests that growth scores either accelerated or decelerated across time, depending on the sign of the quadratic term.

Regardless of the linear or curvilinear shape, though, children's average growth rates were mostly positive between the beginning of pre-kindergarten and the end of first grade. The exception to this trend is the teacher-reported CFBRS construct. Figure 24 shows that CFBRS growth leveled off and became negative across time. This could be explained by the fact that CFBRS scores derive from teacher reports; thus, kindergarten and first grade teachers may have had higher self-regulatory expectations for students than did the pre-kindergarten teachers. Unfortunately, there was no systematic way to test this hypothesis.

Although the indicators mostly indicated average growth across time, six of the seven models also exhibited substantial inter-individual variability in both initial self-regulation status and growth (only DCCS showed non-significant growth rate variance). This significant variability suggests that many children's individual initial self-regulation and growth levels diverged substantially from the sample average.

The inter-individual variance forms the foundation for the remainder of these analyses. The logic is as follows: the models indicate significant variability in self-regulation development; while some of that variance is random, other parts of that variability can likely be explained by observable factors. The present study aims to determine whether Tools and its constituent activities predict variation in children's self-regulation skills. The three sections below, which correspond to research questions two through four, directly address that research objective.

12.2 Research question two: Tools' effect on self-regulation vis-à-vis comparator curricula

This section corresponds to research question two of study two: Does Tools differentially affect children's self-regulation vis-à-vis comparator curricula? While previous studies (Blair & Raver, 2014; Diamond et al., 2007; Lonigan & Phillips, 2012) have investigated this question using randomized designs, none have used structural equation models generally or

latent growth models specifically. That said, the present analysis still addresses a research aim that has been explored in multiple Tools evaluation studies; thus, this particular analysis is best classified as a replication study (Bryman, 2012).

Specifically, this analysis aims to replicate the findings from the previous PRI study (Farran & Wilson, 2014), which used the same dataset as the present study to estimate Tools' effect on children's self-regulation. The present analysis goes beyond that study in three ways. First, this study employs latent variable structural equation modeling to account for the measurement error in the self-regulation construct. By contrast, Farran et al. (2014) used an equally-weighted self-regulation composite variable. In addition to issues of measurement error, PRI's self-regulation composite variable also assumes that each of the five executive function tasks assesses self-regulation with equal weight, which is an unlikely assumption.

Secondly, the present study goes beyond Farran et al. (2014) by investigating the associations between specific Tools activities and children's self-regulation. Farran et al. (2014), as well as all the existing Tools evaluation studies described in this dissertation's systematic review (see Section III), analyzed Tools at the curricular level. Research questions three and four of the present study will be the first to estimate the associations between specific Tools activities and children's self-regulation.

Thirdly, the present study also assesses differences across the two chronologically sequential cohorts (see section 11.2), which was not reported in the previous PRI study using these data (Farran & Wilson, 2014). This sensitivity analysis assesses whether the results vary between the first and second cohorts in the Tools evaluation study. In so doing, the reliability of the results from the Farran & Wilson (2014) paper can be further assessed.

It is important to state that the results for this research question two analysis can be considered causal. This is because the 1140 children in the analytic sample attended schools that were randomly assigned to either Tools or comparator curricula. Thus, if a significant difference between the growth trajectories of Tools students versus comparator students is observed, then that difference can be attributed only to the curriculum they experienced. Consequently, given the cluster-randomized design of this study, the curriculum can be said

to *cause* any observed differences in children's self-regulation outcomes. The findings for each of the three outcome measures are shared in the sections below.

12.2.1 Findings for the latent executive function construct

The latent growth model regressed children's executive function intercept and growth parameters on their curriculum assignment (0 = control, 1 = Tools). These models exhibit two critical differences from the research question one models. First, the models for research questions two through four designate the self-regulation intercept at time point two (i.e., the end of pre-kindergarten, when children had experienced a full school year of Tools). This modification enables analysis of whether Tools predicts changes in children's mean self-regulation scores (i.e., the intercept value) at the end of the implementation year.

If the intercept had been kept at time one (i.e., the beginning of pre-kindergarten), then this would have precluded estimation of Tools' effect on children's early self-regulation skills because Tools implementation would have only just begun. By placing the intercept at time point two, the full impact of Tools after one complete school year can be estimated.

The second contrast from research question one is that the models for research questions two through four specify a latent executive function construct, which comprises the five task-based indicators described in section 11.5.1. That is, instead of modeling each indicator separately as was done in the previous section 12.1, these analyses modeled the five indicators as a singular executive function construct to account for measurement error, which has also been done in other executive function studies (Denham et al., 2012; Wiebe et al., 2011). The explanation for and testing of the executive function construct here can be found in the previous methodology chapter (section 11.6).

After creating the latent executive function construct, the executive function intercept and growth rate were regressed on children's condition assignment (0 = control, 1 = Tools). Once again, this analysis investigates whether Tools causes change in Tools students' executive function intercepts and growth rates vis-à-vis comparison group children's intercepts and growth rates.

The model exhibited acceptable fit ($\chi^2_{[179]} = 632.65$; $p < .001$, RMSEA = .047, CFI = .940, SRMR = .055). The results indicated significant differences between the executive function skills of students in Tools classrooms and students in 'business-as-usual' classrooms. Figure 26 below displays the beta coefficients for the Tools condition as a predictor of children's executive function intercept and slope.

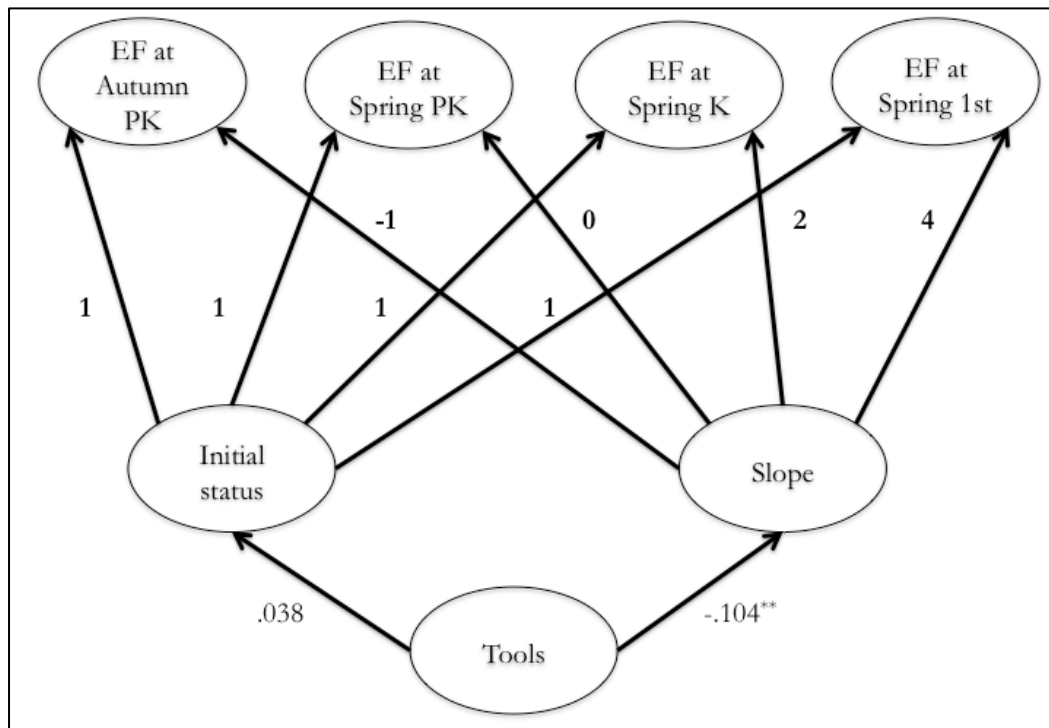


Figure 26: 'Tools' impact on children's initial executive function status and growth

Specifically, figure 26 indicates a significant and negative ($\beta = -.104$, $p < .01$) beta coefficient for the executive function growth parameter. This suggests that children in Tools classrooms exhibited significantly shallower growth trajectories over time than comparison group children.

Thus, these findings replicated the PRI study results (Farran & Wilson, 2014), which also indicated significantly lower composite executive function construct scores by the spring of first grade for Tools children as well as no executive function differences at pre-kindergarten autumn (i.e., the intercept). Although these findings replicate those of Farran & Wilson (2014), the analyses here accounted for measurement error in the executive function

indicators and are based on both student cohorts, thus extending the findings beyond those of Farran & Wilson (2014). The sections below investigate whether differences emerge for teacher-reported and researcher-reported self-regulation (CFBRS and SAR, respectively).

12.2.2 Findings for teacher-reported self-regulation skills

As in the previous model, this model also regressed children's self-regulation intercept and growth parameters on their curricular assignment (0 = control, 1 = Tools). The model exhibited excellent model fit ($\chi^2_{[176]} = 588.156$; $p < .001$, RMSEA = .045, CFI = .975, SRMR = .049), and the results are displayed below in figure 27.

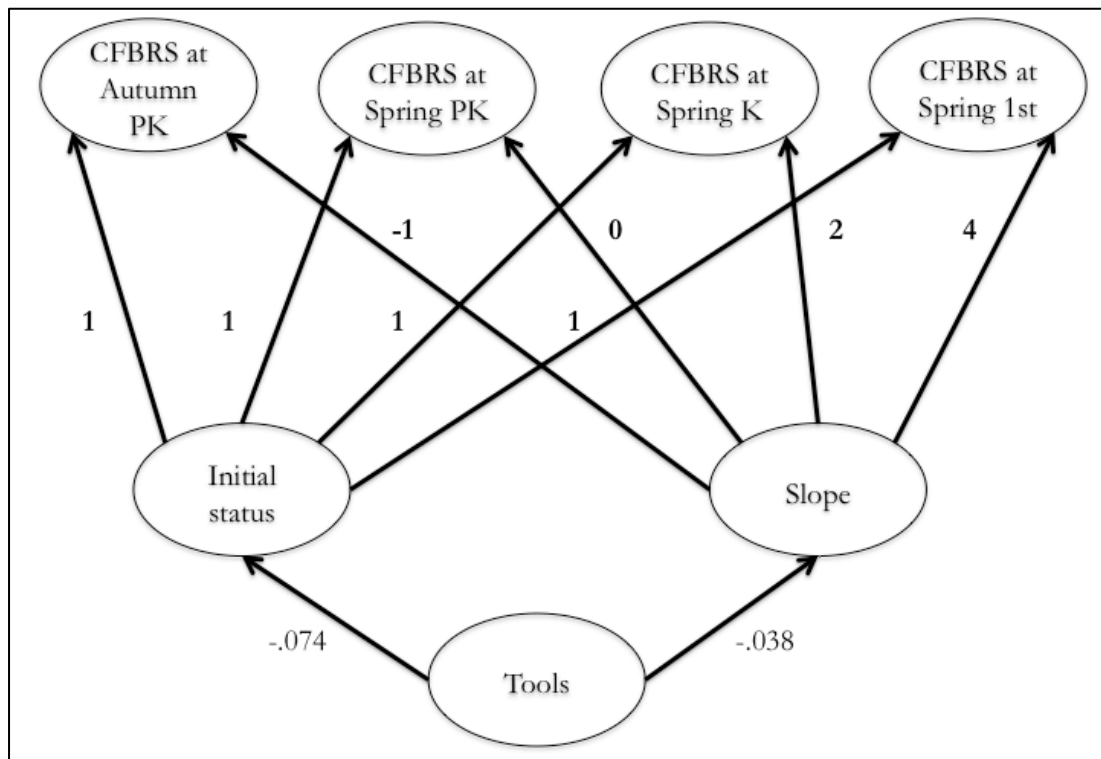


Figure 27: 'Tools' effect on children's initial CFBRS status and growth

The results indicate no significant differences between the intercept or growth rates of CFBRS scores across curricular conditions. The non-significant intercept and growth results replicated PRI's findings of no effect for CFBRS scores (Farran & Wilson, 2014). The section below explores whether any differences are observed for researcher-reported self-regulation (SAR).

12.2.3 Findings for researcher-reported self-regulation skills

Once again, the latent growth model regressed children's self-regulation intercept and growth parameters on their curricular assignment (0 = control, 1 = Tools). The model exhibited excellent model fit ($\chi^2_{[179]} = 542.904$; $p < .001$, RMSEA = .042, CFI = .972, SRMR = .059, and the results are displayed below in figure 28.

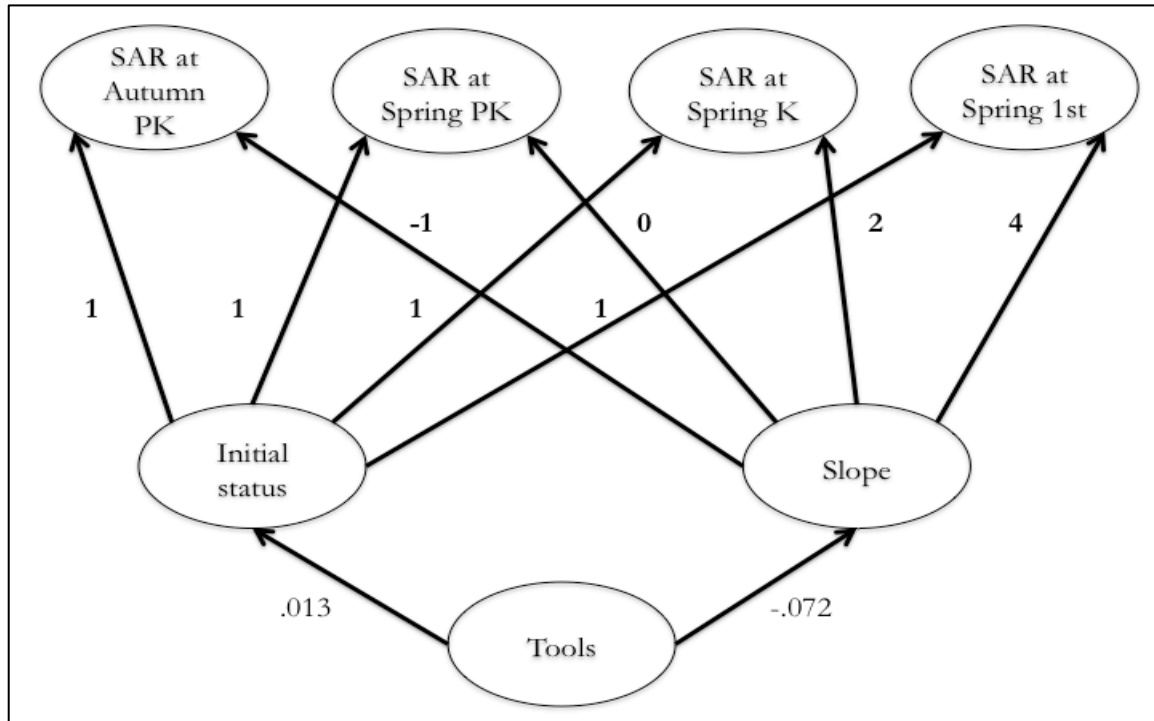


Figure 28: 'Tools' effect on children's initial SAR status and growth

The results indicate no significant differences for the SAR intercept or growth factor. Thus, these findings replicated the results from the PRI study (Farran & Wilson, 2014), which also indicated no differences between Tools and comparator children's SAR scores.

12.2.4 Interaction effect analyses

In addition to the overall effects of Tools on children's self-regulation outcomes, those impacts can also vary by various demographic characteristics. That is, differences in the overall effect of Tools on self-regulation can be broken down by student gender, language learning status (ELL), and special education status (SEN).

For example, even though Tools exhibited no overall effect on children’s teacher-reported self-regulation skills (i.e., CFBRs), it could be the case that Tools predicts higher CFBRs scores for female students specifically, or lower self-regulation for students learning English. Interaction analysis enables assessment of these possibilities. Based on the interaction analysis, none of the observed results varied significantly across student gender, ELL, or SEN status (see table 31).

Table 31: Interaction effects for child gender, ELL, and SEN status

Outcome measure	Gender		ELL		SEN	
	Intercept	Slope	Intercept	Slope	Intercept	Slope
Executive function	.071	-.043	.010	-.010	.023	-.013
CFBRs	.032	.013	-.019	.048	-.054	.075
SAR	.025	.037	.045	-.004	-.004	-.005

12.2.5 Cohort sensitivity analysis

As mentioned in section 11.2, the PRI dataset includes data from two successive cohorts. Cohort one contains 60 classrooms (n = 32 for Tools) and 646 students, whereas cohort two contains 20 classrooms (n = 10 for Tools) and 499 students. PRI selected the two-cohort design to evaluate reliability; that is, were the results of the analysis consistent over two sequential cohort groups?

Thus, for the present research question, it is worthwhile to assess whether Tools differentially affected children’s self-regulatory growth in cohorts one and two. In order to do so, cohort was entered into the latent growth model as a binary grouping variable (1 = child membership in cohort one, whereas 2 = membership in cohort two). The latent growth model then estimates an intercept and growth parameter for each cohort, which enables comparison of the two sets of self-regulation parameter estimates.

The cohort sensitivity models exhibited mostly good fit (RMSEA under .08, CFI above .90). The observed SRMR value of .094 (versus a threshold of .08) for the executive function construct is slightly high, but this should not be a cause for concern given the other fit index

estimates, which all suggest adequate fit (Bryne, 2012). Table 32 below displays the intercept and growth rate parameter estimates for each cohort sensitivity analysis model corresponding to each outcome measure.

Table 32: Cohort sensitivity analysis results across all three outcome measures

Outcome measure	Cohort 1		Cohort 2	
	Intercept	Slope	Intercept	Slope
Executive function	.053	-.234***	.162	-.140
CFBRS	-.063 ⁺	.039	.063	-.201**
SAR	-.024	.005	-.034	.036

Although the results were not sensitive to cohort assignment for SAR, the results did vary by cohort for the CFBRS and executive function construct. For the CFBRS, Tools students in cohort one exhibited a marginally lower self-regulation intercept ($\beta = -.063$, $p = .086$) at the end of pre-kindergarten; for cohort two, Tools students exhibited significantly shallower growth trajectories ($\beta = -.201$, $p < .01$) than control group children. For the executive function construct, Tools cohort one children exhibited significantly shallower growth trajectories ($\beta = -.234$, $p < .001$) than their control group counterparts.

Overall, the results were only stable between cohorts one and two for the SAR construct. Negative effects were observed in both cohort one (for the executive function construct) and cohort two (for the CFBRS ratings). Thus, we cannot say that one Tools cohort performed better than the other since Tools students in both cohorts exhibited significantly lower self-regulation scores than the comparison group on one of the outcome measures. In sum, not only did these results reveal the presence of cohort effects, but they also further corroborated the finding that Tools children exhibited lower self-regulation scores than did control group children.

12.2.6 Summary of research question two

The findings from the latent growth analyses mostly replicated those of the PRI study (Farran & Wilson, 2014). In line with Farran et al. (2014), Tools children exhibited

significantly lower executive function skills across time than did control group children, whereas no significant effects were observed for researcher- and teacher-reported self-regulation abilities. Going beyond the Farran et al. (2014) study, the cohort sensitivity analysis revealed negative effects for Tools students in both cohort one and two.

Although the present study went beyond the previous PRI study, it did use the same dataset to address a similar question: does Tools differentially affect children's self-regulation skills vis-à-vis comparison curricula? By contrast, the next two sections below present results for two previously unaddressed research questions: Firstly, which blocks of Tools activities predict children's self-regulation developmental trajectories? And, secondly, which specific Tools activities predict children's self-regulation developmental trajectories? Neither of these questions has yet been addressed in the existing literature.

12.3 Research question three: Subject-specific Tools activity blocks and self-regulation

Although several existing studies have analyzed Tools' effectiveness at the curricular level (Blair & Raver, 2014; Diamond et al., 2007; Lonigan & Phillips, 2012; Morris et al., 2014), no research has yet examined the associations between specific Tools' activities and children's self-regulation. Research questions three and four addressed that precise issue. Since comparison group classrooms did not use Tools activities, which are unique to Tools, the analytic sample for this section and the next was comprised only of Tools children (n=646).

Once again, the present analysis for research question three investigated Tools activity groupings and their associations with children's self-regulation development. For this analysis, I used the Tools manual (Leong & Bodrova, 2011) to determine appropriate groupings of Tools activities. The manual includes clearly delineated sections for different subjects and enumerates a set of activities that fit into each subject designation. For example, one section entitled "Literacy" (Leong & Bodrova, 2011, p. 247-413) includes 14 activities that target children's foundational literacy skills.

Thus, this analysis aimed to investigate the Tools activity blocks as outlined by the Tools developers and manual. It also would have been possible to design my own activity

groupings and tested the curriculum in an ad hoc method; instead, however, I chose to test the curriculum as prescribed by the manual in order to maximize the usefulness of this study for educators.

In order to create activity block variables, I averaged the activities from each block in the Tools manual into a single composite score (see Appendix J for the full list of activities for each grouping). Table 33 below shows each activity block's name as well as the descriptive statistics for that activity block's implementation level across teachers. Table 33 also shows mostly high internal consistency values for each activity block, which underscores the appropriateness of creating a subject-specific composite variable (e.g., averaging all the Tools activities into a single 'math' composite variable).

Once again, the main analyses in this section used the teacher-reported activity implementation values (see section 11.7.1 for a review of this variable). Thus, a score of 0 signifies that the teacher never implements activities from the block, whereas a value of 3 signifies that the teacher implements the activities daily.

As table 33 indicates, Tools teachers engaged the most in make-believe play followed by the introduction activities and literacy. Make-believe play also had the highest levels of dispersion, which indicates that different teachers implemented play at different frequencies. Teachers reported nearly identical implementation frequencies and dispersion levels for attention-focusing activities, math, and science.

Science does not have a Cronbach α value because the Tools manual contains only one activity (again, see Appendix J for the full set of activities in each activity block). Consequently, there is no way to calculate internal consistency among multiple items if only one science item exists. Nonetheless, science was included as a block in the analyses because science is represented as its own section in the manual. Thus, in order to ensure consistency in my analytic approach and test the manual as it is designed, science was included as its own activity block.

Table 33: Teacher-reported implementation frequency descriptive statistics for the Tools activity blocks

Activity block	Min	Max	Mean	SD	Cronbach α
Attention-focusing	0.80	2.65	1.77	0.19	0.82
Introduction	1.78	2.89	2.43	0.06	0.61
Literacy	1.00	2.93	1.97	0.20	0.80
Make-believe play	1.00	3.00	2.59	0.26	0.85
Math	0.93	2.64	1.78	0.17	0.89
Science	0.00	2.50	1.75	0.19	N/A

Although some groupings have generic titles such as literacy and math, it is key to restate that the activities in those groups are Tools-specific and inspired by the Vygotskian tradition. In other words, the individual activities within the literacy, math, and other groupings all aim to incorporate both an academic and self-regulatory component. Thus, the activity groupings have conventional names but consist of relatively unconventional activities.

Another relevant note is that this analysis involves the teacher-reported implementation frequency data as a predictor of self-regulation. By contrast, the researcher-observed fidelity data (section 11.7.2) was used as a sensitivity analysis (see section 12.3.7). The implementation frequency data were used for the main analysis because they were considered more reliable than the fidelity data, which was described in the previous methodology chapter (section 11.10.5).

The final relevant note is that the activity analyses relate only to Tools students because only Tools students experienced Tools activities. Thus, as described in section 11.9.2, the sample size for the remainder of the analyses is the 646 children in Tools classrooms (i.e., not the 1140 children from the full sample of both Tools and non-Tools classrooms).

Now, the sections below share results for the association between children's self-regulation skills and each activity block: make-believe play, literacy, math, science, attention-focusing activities, and introduction.

12.3.1 Make-believe play block

Hypothesis 5 anticipates positive associations between the make-believe play activity block and children's self-regulation. According to the curricular developers, "the best way for

preschool and kindergarten children to practice self-regulatory behaviors is to engage in mature make-believe play” (Bodrova & Leong, 2015b). The make-believe play block consists of four activities: 1) planning for the play scenario with other children and teachers, 2) practicing the scenario with teacher support, 3) engaging in the play scenario with minimal teacher support, and 4) cleaning up at the end of the make-believe play session (Leong & Bodrova, 2011).

Although the make-believe activity block technically contains four activities, only three were included in this analysis. Make-believe play cleanup was omitted because it was highly collinear ($r = .94$) with play center time (i.e., the third activity in the paragraph above). That is, nearly every time students engaged in make-believe play centers, they also engaged in play center cleanup. By contrast, play planning and play practice occurred only when children were starting a new scenario or theme, which happened more sparingly relative to the daily implementation of play centers and cleanup.¹⁹

Tabachnick and Fidell (2013) recommend dropping one of the collinear items from the model in order to obtain unbiased parameter estimates. Given Tools’ focus on make-believe play as the critical driver of self-regulation growth (Bodrova & Leong, 2007), the play centers were included in the activity block composite instead of the play cleanup, which lacks a strong theoretical linkage to promoting children’s self-regulation growth.

All models for the make-believe play activity block exhibited acceptable fit (RMSEA and SRMR under .08, CFI above .90). Table 34 depicts the standardized beta coefficients of the make-believe play block as a predictor of the intercept and growth rate for the researcher-reported self-regulation (SAR), teacher-reported self-regulation (CFBRS), and executive function constructs.

¹⁹ Both play planning and practice exhibited correlations below the $r = .9$ threshold recommended by Field (2013). The tolerance and variance inflation factor (VIF) estimates, which assess multi-collinearity, were also all within Field’s (2013) recommended thresholds.

Table 34: Beta coefficients for the make-believe play (MBP) block on all three outcome measures

	MBP → Intercept	MBP → Slope
Executive function	.080	-.094 ⁺
CFBRS	.012	-.083
SAR	-.141 ^{**}	.046

In contrast to Hypothesis 5, the make-believe play block exhibited non-significant associations with the teacher-reported CFBRS. Moreover, the make-believe play activity block was associated with a significantly lower researcher-reported SAR intercept at the end of pre-kindergarten (i.e., the end of the Tools implementation year) and a marginally lower executive function intercept. Both of those outcomes contravene Hypothesis 5 of a positive association between the make-believe play block and self-regulation.

12.3.2 Literacy block

Although research indicates that self-regulation skills support children's literacy learning (Blair & Raver, 2015; McDonald et al., 2011), recent research (Fuhs et al., 2015) shows no reciprocal effect of children's literacy skills improving their self-regulation. Once again, Tools literacy activities such as buddy reading were designed to target self-regulation, but prior evidence of literacy activities' promotion of self-regulation skills is lacking. Thus, Hypothesis 5 does not anticipate a significant association between the Tools literacy activity block and children's self-regulation skills.

All models for the literacy activity block exhibited acceptable fit (RMSEA and SRMR under .08, CFI above .90). Table 35 depicts the intercept and growth rate estimates for researcher-reported self-regulation (SAR), teacher-reported self-regulation (CFBRS), and the executive function construct. The results are consistent with Hypothesis 5 in that there were no significant associations between the Tools literacy block and children's self-regulation intercept or growth rate across any of the three outcome measures.

Table 35: Beta coefficients for the literacy activity block on all three outcome measures

	Lit → Intercept	Lit → Slope
Executive function	-.051	-.077
CFBRS	-.075	.040
SAR	-.075	-.083

12.3.3 Math

Hypothesis 5 anticipates significant and positive associations between math activities and self-regulation given the strong connection observed in the existing literature (Bull & Scerif, 2010; Clements & Sarama, 2012; Gilmore et al., 2013). Not only do math activities in general support self-regulation growth (Fuhs et al., 2015), but evidence of self-regulation promotion through math activities also exists in the literature (Clements et al., 2016), and Tools math activities are especially designed to target children’s self-regulation.

For example, in the “making collections” activity (Leong & Bodrova, 2011, p. 441), children work in pairs to create sets of counters. One child makes the sets while the other child serves as the checker to ensure that the set is correct; the children then switch roles at the end (Leong & Bodrova, 2011, p. 441). Similar to buddy reading, this activity requires children to remember their roles and associated tasks, flexibly switch across roles at the appropriate times, and inhibit the impulse to switch roles at inappropriate times.

All models for the math activity block exhibited acceptable fit (RMSEA and SRMR under .08, CFI above .90). Table 36 depicts the intercept and growth rate estimates for researcher-reported self-regulation (SAR), teacher-reported self-regulation (CFBRS), and the executive function construct. In contrast to Hypothesis 5, the executive function intercept was marginally significant and negative ($\beta = -.103$, $p = .091$). No other significant associations were observed.

Table 36: Beta coefficients for the math activity block on all three outcome measures

	Math → Intercept	Math → Slope
Executive function	-.103 ⁺	-.039
CFBRS	-.102	.083
SAR	-.037	-.094

12.3.4 Science

The Tools manual includes only one science activity called “science eyes” (Leong & Bodrova, 2011, p. 502). Nonetheless, the manual does explicitly include a science section; thus, for the sake of consistency, the Tools science block is tested here even though it consists of only one activity. Specifically, science eyes requires pairs of students to observe objects using a magnifying glass and then report on what they see (Leong & Bodrova, 2011, p. 502). As with many other Tools activities, the partner pairings require the children to assume and flexibly switch between roles at appropriate times.

Although self-regulation skills have been shown (Schraw et al., 2006; Velayutham, Aldridge, & Fraser, 2011) to predict improved science learning, the literature lacks evidence that science learning reciprocally improves self-regulation. Thus, Hypothesis 5 does not anticipate an association between the science eyes activity and children’s self-regulation intercept or growth rate.

The science activity block models exhibited acceptable fit (RMSEA and SRMR under .08, CFI above .90). Table 37 depicts the intercept and growth rate estimates for researcher-reported self-regulation (SAR), teacher-reported self-regulation (CFBRS), and the executive function construct. The results are consistent with Hypothesis 5 in that there were no significant associations between the Tools science block (i.e., science eyes) and children’s self-regulation intercept or growth rate across any of the three outcome measures.

Table 37: Beta coefficients for the science block on all three outcome measures

	Science → Intercept	Science → Slope
Executive function	-.059	-.021
CFBRS	-.056	-.016
SAR	-.077	.039

12.3.5 Attention-focusing activities

Hypothesis 5 anticipates that attention-focusing activities will be positively associated with children's self-regulation development. This hypothesis derived from the notion that these activities were especially designed to tax children's self-regulation (Leong & Bodrova, 2011). For example, activities such as freeze dance require children to remember a dance pose, inhibit the impulse to strike the pose until the music stops, and then switch quickly across different poses as the teacher changes visual cards. Because each of the 10 activities in this attention-focusing group aims to target children's executive function, this block was hypothesized to be associated with higher self-regulation intercept and growth values.

All models for the attention-focusing activity block exhibited acceptable fit (RMSEA and SRMR under .08, CFI above .90). Table 38 depicts the intercept and growth rate estimates for researcher-reported self-regulation (SAR), teacher-reported self-regulation (CFBRS), and the executive function construct.

Table 38: Beta coefficients for the attention-focusing activity block on all three outcome measures

	Attention → Intercept	Attention → Slope
Executive function	.028	-.073
CFBRS	.033	-.125
SAR	-.106 ⁺	-.032

Contrary to Hypothesis 5, table 38 indicates that attention focusing-activities are not associated with significant change in the executive function or CFBRS constructs. Moreover, attention-focusing activities are associated with a marginally lower SAR intercept ($\beta = -.106$, $p = .097$), which contravenes Hypothesis 5.

12.3.6 Introduction activities

Hypothesis 5 did not anticipate significant associations between the introduction activity block and children's self-regulation. Although the Tools manual asserts that all activities contain a self-regulatory component, the introduction activities most resemble traditional classroom activities observed across preschool curricula. For example, the introduction block contains activities such as a daily calendar routine, a weather check, unspecified community building activities, name games, and others (again, the full list of activities in each group is presented in Appendix J).

All models for the introduction activity block exhibited acceptable fit (RMSEA and SRMR under .08, CFI above .90). Table 39 depicts the intercept and growth rate estimates for researcher-reported self-regulation (SAR), teacher-reported self-regulation (CFBRS), and the executive function construct. In line with Hypothesis 5, the introduction activity grouping lacked significant associations with both the self-regulation intercept and growth rate across all outcomes.

Table 39: Beta coefficients for the introduction activity block on all three outcome measures

	Intro → Intercept	Intro → Slope
Executive function	-.069	-.030
CFBRS	-.097	.060
SAR	-.074	-.046

12.3.7 Sensitivity analysis with researcher-observed implementation fidelity

As described in section 11.7, Tools activity implementation data was collected in the form of both teacher-reported and researcher-observed measures. In the present study, the teacher-report measures served as the main measure for analysis because the PRI researchers only observed Tools classrooms for three days over the course of the year (Farran, Wilson, Meador, Norvell, & Nesbitt, 2015). Thus, the researcher-observed measure may be less representative of teachers' actual activity implementation throughout the year than the teacher-reported measure (Meador, 2015).

Although the teacher-report measure was selected as the main measure of analysis, the researcher-observed data was still used for sensitivity analysis. That is, are the results consistent across both Tools implementation measures? In order to conduct this analysis, the same activity grouping blocks were used from the Tools manual (Leong & Bodrova, 2011) that were used in the previous analysis. The results across all blocks are indicated below in table 40. In table 40, each numerical cell indicates the beta coefficient of the latent self-regulation intercept and slope ('I' and 'S,' respectively) regressed on the activity block.

Table 40: Sensitivity analysis with researcher-observed fidelity data

	Executive function		SAR		CFBRS	
	I	S	I	S	I	S
Attention-focusing	.04	-.10	-.10	-.01	.05	-.12
Introduction	.02	-.17*	-.15**	-.02	-.02	-.07
Literacy	.01	-.08	-.12 ⁺	.065	-.02	-.05
Make-believe play	-.04	-.09	-.17*	.21*	-.04	.02
Math	.05	.03	-.03	.07	-.01	.03
Science	.09	-.08	-.08	.11	.04	.09

The values in table 40 demonstrate broad consistency with the findings from sections 12.3.1 through 12.3.6. Specifically, none of the activity blocks predicted exclusively positive associations with self-regulation development. That is, if a significantly positive association was observed for the intercept, then it was counterbalanced by a negative association on the slope or vice versa. The results of each block are summarized in the bulleted list below.

- **Attention-focusing block:** Whereas the implementation frequency analysis (see section 12.3.5) exhibited a marginally significant negative association ($\beta = -.11$, $p = .09$) with the teacher-reported SAR intercept, the current sensitivity analysis with implementation fidelity data yielded no significant associations with any outcomes.
- **Introduction block:** Whereas the introduction activities exhibited no associations in the implementation frequency analysis (see section 12.3.6), this sensitivity analysis with the fidelity data yielded negative associations with both the executive function slope ($\beta = -.17$, $p < .05$) and SAR intercept ($-.15$, $p < .01$).

- **Literacy block:** Whereas the literacy block exhibited no associations in the implementation frequency analysis (section 12.3.2), this sensitivity analysis with the implementation fidelity data exhibited a marginally negative ($\beta = -.11$, $p = .082$) association with the researcher-reported SAR intercept.
- **Make-believe play block:** Whereas the make-believe play block exhibited a negative association with the researcher-reported SAR in the implementation frequency analysis (see section 12.3.1), this sensitivity analysis with the fidelity data also exhibited a negative association ($\beta = -.17$, $p < .05$) with the SAR intercept but a positive association ($\beta = .21$, $p < .05$) with the SAR slope.
- **Math block:** Whereas the math block exhibited a marginally significant negative ($\beta = -.10$, $p = .09$) association with the executive function intercept in the implementation frequency analysis (see section 12.3.3), this sensitivity analysis with fidelity data indicated no significant associations with any outcomes.
- **Science block:** The implementation frequency analysis with the science block (see section 12.3.4) indicated no associations with any outcomes, and this sensitivity analysis also indicated no associations with any outcomes.

Once again, the sensitivity analysis using the implementation fidelity data generally corroborated the results based on the implementation frequency data. Most importantly, as with the implementation frequency data analyses, none of the activity blocks in these sensitivity analyses predicted consistently higher self-regulation skills.

12.3.8 Summary of Tools activity blocks and children's self-regulation

The results did not support Hypothesis 5 regarding the associations between Tools activity blocks and children's self-regulation. In fact, the negative associations observed among at least one of the three outcome measures for attention-focusing activities, make-believe play, and math activity blocks directly contravened Hypothesis 5.

Although informative, these activity block analyses do not reveal which specific activities within the groupings account for the observed associations. Thus, the upcoming, and final, results section disaggregates the make-believe play, math, and attention-focusing activity

blocks into their constituent activities to identify which specific activities predict self-regulation and in which direction. Given that only those three blocks significantly predicted children's self-regulation skills, only the activities in these three blocks will be assessed in the following section (the results for all remaining Tools activities are also presented in Appendix K).

12.4 Research question four: Specific Tools activities and self-regulation

Tools includes 61 discrete instructional activities, each of which is intended to target children's self-regulation skills (Bodrova & Leong, 2007). To the best of my knowledge, no existing study has investigated the associations between specific Tools activities and children's self-regulation. Table 41 below provides the zero-order correlations between each teacher-reported Tools activity and the three self-regulation outcomes at the end of Tools implementation (i.e., pre-kindergarten spring). These correlations do not control for the vector of covariates described in section 11.8, whereas the sections below will use latent growth models to isolate the unique association of the activity with self-regulation.

Table 41: Raw correlations among each Tools activity and the three outcome measures

Activity name	Executive Function	SAR	CFBRS
Attention-focusing	-.041	-.102**	-.058
Attribute game	-.039	-.105**	-.055
Buddy reading	-.004	-.073	-.046
Community building	-.136**	-.04	-.089*
Complete and continue	-.06	-.028	-.128**
Elkonin boxes	.025	-.091*	-.097*
Freeze game	-.091*	-.056	-.039
Freeze on number	-.076	-.043	-.091*
Graphics practice	-.080*	-.077*	-.142**
I have colors	-.110**	-.080*	-.170**
I have letters	-.074	-.100*	-.130**
I have names	-.083*	-.06	-.174**

I have numbers	-.089*	-.111**	-.132**
I have shapes	-.106**	-.047	-.148**
Introduction to centers	.007	-.159**	-.085*
Making collections	-.032	-.049	-.073
Make a rhyme	-.038	-.122**	-.085*
Math memory	.019	-.058	-.035
Make believe play centers	-.008	-.108**	-.141**
Make believe play cleanup	.032	-.085*	-.079*
Make believe play planning	.100*	-.043	-.006
Make believe play practice	.106**	-.131**	.05
Message of the day	-.064	-.003	-.04
Mouse trap	.061	-.084*	.148**
Mr. Wolf	.108**	-.055	.118**
Mystery activities	.03	.069	.145**
Number follow the leader	-.039	-.091*	.031
Numeral games	-.091*	-.079*	-.128**
Numberline hopscotch	-.090*	-.158**	-.134**
Partner freeze	.019	-.175**	-.05
Patterns and manipulatives	-.091*	-.101*	-.181**
Pattern movement	.029	-.073	-.087*
Pretend transitions	-.055	-.109**	.039
Puzzles with manipulatives	.025	-.053	.036
Remember and replicate	-.080*	-.048	-.058
Science eyes	.008	-.080*	-.045
Share the news	-.014	-.033	.014
Story lab activities	.008	-.072	-.011
Takeaway sounds	.006	-.115**	-.065
Tally	-.006	-.095*	-.026
Timeline calendar	-.051	-.03	-.087*
Two step freeze	.053	-.096*	.059
Venger drawings	-.093*	-.153**	-.132**

Weather graph	-.008	.015	.028
Write a fingerplay	.055	-.073	-.015

Although table 41 depicts a host of significant correlations, not all activity results will be presented in the main text body. Instead, only the constituent activities from the significant Tools activity blocks (i.e., make-believe play, math, and attention-focusing blocks) will be shared here. Results from all other activities are documented in Appendix K. The constituent activities from each of the three activity blocks are briefly described in the sections below,²⁰ and their associations with the self-regulation intercept and growth rates will also be reported.

12.4.1 Make-believe play activities

The make-believe play block contains four activities: 1) play planning, 2) play practice, 3) play centers, and 4) play cleanup. Once again, the fourth activity, make-believe play cleanup, was omitted from this analysis because it was highly collinear with the make-believe play activity ($r = .94$). Thus, the sections below share results from each of the three play activities and their individual associations with children's self-regulation. All models for the make-believe play activities exhibited acceptable fit (RMSEA and SRMR under .08, CFI above .90).

Make-believe play planning

This activity requires a small group of children to work with a teacher to plan a play scenario (e.g., people at a grocery store). Once the scenario has been determined, children negotiate roles (e.g., one child is a grocer, one is a customer, one is a manager, and one is a butcher). The children then write make-believe play plans; depending on the children's writing level, they may simply draw a picture of their roles or write a description using words and pictures.

²⁰ The activity descriptions also could have been shared in the methodology section, which may seem like a more conventional approach to research reporting. However, because I did not know which activity blocks would yield significant associations with children's self-regulation, I did not know which activities would be analyzed in this section. Instead, I had to analyze the data first as presented in the results section 12.3 to determine which activities to report in the present section. Thus, the activity descriptions are found in this results section as opposed to the methodology section.

The play plans then serve as mediators for children to remember and successfully act out their roles in the play scenario (Leong & Bodrova, 2011, p. 152).

Three latent growth models were specified, with one model per outcome measure. Each model regressed children’s self-regulation intercept and growth rate on the make-believe play planning activity (while controlling for the vector of covariates). Table 42 shows the beta coefficient of the play planning activity on the intercept and slope of each outcome measure.

The play planning activity, and each of the activities for the research question four analyses, are labeled as “Act” (shorthand for “Activity) in the tables due to spatial considerations. Table 42 below indicates marginally significant and negative associations with the executive function slope as well as the SAR intercept, but no conventionally significant associations with any of the self-regulation outcomes were observed.

Table 42: Beta coefficients for make-believe play planning (Act) on the three outcome measures

	Act → Intercept	Act → Slope
Executive function	-.089	-.087 ⁺
CFBRS	.032	-.099
SAR	-.087 ⁺	.043

Make-believe play practice

Make-believe play practice involves both teachers and children collaboratively practicing for the make-believe play scenario. Specifically, the teacher models appropriate role behavior for a target scenario, and the children pretend alongside the teacher. Thus, the teacher scaffolds the child’s play abilities, which is especially important at the beginning of the school year and when “teachers are introducing or expanding new themes” (Leong & Bodrova, 2011, p. 173).

Once again, the play practice activity, and each of the activities for the research question four analyses, is labeled as “Act” in the tables due to spatial considerations. Table 43 below indicates a negative and significant association ($\beta = -.167$, $p < .001$) between play practice and the SAR construct. Thus, at the end of Tools implementation in the pre-kindergarten

year (i.e., the designated intercept), children whose teachers reported higher play practice implementation exhibited lower researcher-reported self-regulation skills. This result mirrors the findings from the make-believe activity block analysis, which also indicated a significantly negative association between the make-believe play block and SAR intercept. No other associations with other measures' intercepts or growth rates were observed.

Table 43: Beta coefficients for make-believe play practice on the three outcome measures

	Act → Intercept	Act → Slope
Executive function	.085	-.071
CFBRS	.051	-.078
SAR	-.167***	.057

Make-believe play centers

After the children have planned and practiced their play, make-believe play centers allow children to act out their play scenario. Tools play scenarios are intended to involve minimal teacher intervention; if children deviate from their plan, then the play plan mediators should help them recover their focus. Tools play scenarios are also intended to be prolonged, uninterrupted blocks that last between 45 minutes and an hour per day (Leong & Bodrova, 2011, p. 186). Once again, the Tools developers consider make-believe play to be the most powerful driver of self-regulation growth (Bodrova & Leong, 2007).

Table 44 below indicates that play centers predicted a significantly lower intercept for the SAR construct ($\beta = -.119$, $p < .05$). That standardized beta coefficient is slightly smaller in magnitude than that of play practice, but both activities exhibited negative associations.

Table 44: Beta coefficients for the make-believe play centers on the three outcome measures

	Act → Intercept	Act → Slope
Executive function	.012	-.089
CFBRS	-.094	-.024
SAR	-.119*	.009

Thus, at the end of Tools implementation in the pre-kindergarten year, children whose teachers reported higher make-believe play implementation exhibited lower researcher-reported self-regulation skills. As with play practice, this result also mirrors the findings regarding the significantly negative association between the make-believe play activity block and the SAR intercept. No other associations with other measures' intercepts or growth rates were observed.

Make-believe play activities summary

Results from the sections above indicate that two of the three make-believe play activities predicted lower self-regulation outcomes. Specifically, play practice and play centers both predicted significantly lower researcher-reported intercepts at the end of pre-kindergarten. This analysis goes beyond the make-believe play block analysis by disaggregating the play block activities and illuminating which specific activities explain the negative association with SAR at the activity block level. In this case, play practice and play centers emerged as the two activities that explain the negative association observed at the group level.

12.4.2 Attention-focusing activities

The sections below share results from each of the ten attention-focusing activities and their individual associations with children's self-regulation. As with the make-believe play activity models, all attention-focusing activity models exhibited acceptable fit (RMSEA and SRMR under .08, CFI above .90). Unlike the make-believe play block, none of the ten attention-focusing activities exhibited collinearity issues, so all ten are analyzed in the sections below.

Freeze game

The freeze game activity requires children to imitate a body configuration depicted on a card that a teacher holds aloft during a class dance break. The children must wait until the music stops to strike the pose, which they must remember from the card, which the teacher has only momentarily held up as children are dancing to a song. Thus, the children have to 1) remember and act out the card's content after it has been removed from view, 2) switch body poses without confusion across each card, and 3) inhibit the impulse to strike the card's pose when they first see the card and instead wait until the music stops.

Table 45 below indicates that freeze dance predicted a significantly lower intercept for the executive function construct ($\beta = -.13$, $p < .05$) and SAR ($\beta = -.122$, $p < .05$). Thus, at the end of Tools implementation in the pre-kindergarten year (i.e., the designated intercept), children whose teachers reported higher freeze game implementation exhibited both lower task-based and researcher-reported self-regulation skills. No other associations with the CFBRs intercept or any growth rates were observed.

Table 45: Beta-coefficients for freeze game on the three outcome measures

	Act → Intercept	Act → Slope
Executive function	-.130*	.018
CFBRs	-.096	-.013
SAR	-.122*	.091

Freeze on number

Freeze on number is similar to freeze game in that children dance to a song and then have to perform an action when the music stops. In freeze on number, the teacher momentarily holds up a card with some number of objects on it, and then the teacher removes the picture from children's view as the music continues. When the music stops, children freeze and hold up the correct number on their fingers before the music starts again (Leong & Bodrova, 2011, p. 34).

Although this activity is structurally similar to freeze game, it exhibited different associations with children's self-regulation development (see table 46). Specifically, freeze on number predicts a significantly lower ($\beta = -.121$, $p < .05$) CFBRs intercept. That is, children whose teachers reported higher freeze on number implementation exhibited significantly lower teacher-reported self-regulation intercepts. No other associations were observed.

Table 46: Beta coefficients for Freeze on Number on the three outcome measures

	Act → Intercept	Act → Slope
Executive function	-.086	-.054
CFBRS	-.121*	-.015
SAR	-.049	-.022

Partner freeze

Partner freeze is the same activity as freeze dance except that the cards help aloft by the teacher depict body positions for two children. That is, two children are meant to strike the appropriate pose in pairs (e.g., a card directs the partners to join hands above their heads to make a human tower).

Although this activity is structurally similar to freeze game, it exhibited different associations with children's self-regulation development (see table 47). Specifically, it does not predict intercepts or growth rates for any outcome measures except the executive function growth rate, where it has a marginally significant negative association ($\beta = -.114$, $p = .057$). Thus, children whose teachers reported higher partner freeze implementation exhibited marginally lower executive function growth rates, but no other associations were significant.

Table 47: Beta coefficients for partner freeze on the three outcome measures

	Act → Intercept	Act → Slope
Executive function	.085	-.114 ⁺
CFBRS	.012	-.045
SAR	-.066	-.123

Two-step freeze

Two-step freeze increases the complexity of freeze dance by including two freeze poses that are color-coordinated. Children briefly view both body pose cards as the music plays. The cards are then removed by the teacher, who holds up a color card, which corresponds to only one of the two body position cards. When the music stops, children must strike the appropriate body pose that corresponds to the color card previously held aloft by the teacher

(Leong & Bodrova, 2011, p. 35). Despite this ostensibly challenging self-regulatory task, two-step freeze implementation did not significantly or marginally predict any self-regulation intercepts or growth rates (see table 48 below).

Table 48: Beta coefficients for two-step freeze on the three outcome measures

	Act → Intercept	Act → Slope
Executive function	.062	-.044
CFBRS	.048	-.060
SAR	-.076	-.049

Pattern movement game

Pattern movement requires children to physically act out patterns provided by the teacher. For example, a teacher may present a pattern that alternates between circles and squares. The teacher instructs the children to touch their knees when they see a circle and their shoulders when they see a square. Teachers can then reverse the association (i.e., touch shoulders with a circle and knees with a square), or they can add objects to the pattern to enhance its complexity (Leong & Bodrova, 2011, p. 37-38). The pattern movement game did not significantly or marginally predict any self-regulation intercepts or growth rates (see table 49 below).

Table 49: Beta coefficients for the pattern movement game on the three outcome measures

	Act → Intercept	Act → Slope
Executive function	.043	-.057
CFBRS	-.012	-.005
SAR	-.044	-.057

Number follow the leader

Number follow the leader requires children to perform a pre-specified action a pre-specified number of times. The teacher selects an action and a number (e.g., jump five times), and then the children act it out. As children master the activity, more complex actions are chosen, and students take over the process of selecting actions and cards. Number follow

the leader did not significantly or marginally predict any self-regulation intercepts or growth rates (see table 50 below).

Table 50: Beta coefficients for number follow the leader on the three outcome measures

	Act → Intercept	Act → Slope
Executive function	.006	-.082
CFBRS	.035	-.112
SAR	-.034	-.093

Mouse trap

Mouse trap is a game where some children are designated as mice and others are designated as mouse traps. “Mice” pretend to eat cheese inside the “trap,” which is constituted by the group of trap children with their arms raised. When the music stops, the trap children lower their arms to catch the mice children, who must run away before being ensnared. According to the Tools manual, “children who are the trap have to inhibit the desire to lower their arms and close the trap until the appropriate moment. Both groups of children first have to inhibit a specific physical behavior and then initiate a specific behavior” (Leong & Bodrova, 2011, p. 41).

Mouse trap predicted a significantly higher intercept ($\beta = .145$, $p < .05$) and a significantly lower slope ($\beta = -.232$, $p < .01$) for CFBRS (see table 51). Thus, at the end of Tools implementation in the pre-kindergarten year, children whose teachers reported higher mouse trap implementation exhibited higher teacher-reported self-regulation, but those children then slowed down relative to their peers over time through a shallower growth rate. No other associations were observed.

Table 51: Beta coefficients for mouse trap on the three outcome measures

	Act → Intercept	Act → Slope
Executive function	.079	-.022
CFBRS	.145*	-.232**
SAR	-.037	-.071

What are you doing, Mr. Wolf?

This activity designates one child as the wolf, whereas the other children receive no distinctive role. The children slowly advance toward the wolf and ask him or her questions. When the wolf eventually answers with, “I’m brushing my teeth,” this signifies that the children must run to a pre-designated safe zone before the wolf catches them (Leong & Bodrova, 2011, p. 42). The class can then shift to different answers besides tooth-brushing to signify the wolf’s attack, which requires children to remember the target phrase and flexibly switch across cue phrases.

As with mouse trap, the wolf activity predicts a significantly higher intercept ($\beta = .094$, $p < .05$) and significantly lower slope ($\beta = -.185$, $p < .05$) for CFBRs. Thus, at the end of Tools implementation in the pre-kindergarten year, children whose teachers reported higher wolf activity implementation exhibited higher teacher-reported self-regulation, but those children then slowed down relative to their peers over time through a shallower growth rate. No other associations were observed (see table 52).

Table 52: Beta coefficients for the wolf activity on the three outcome measures

	Act → Intercept	Act → Slope
Executive function	.094	-.016
CFBRs	.142*	-.185*
SAR	-.062	-.021

Pretend transition

Pretend transition aims to harness Vygotsky’s focus on private speech into the classroom context (Leong & Bodrova, 2011, p. 50). Specifically, when transitioning between activities, the teacher directs children to pretend to be some other organism while using private speech (e.g., flap your hands like a bird and whisper ‘flap’ to yourself). In this activity, children’s intentional use of private speech is meant to focus their attention on a specific role, much like the make-believe play scenarios. The act of inhibiting one’s own tendencies to walk like themselves, and instead guiding their actions through private speech, aims to promote self-regulation abilities (Leong & Bodrova, 2011, p. 50).

Table 53 indicates that pretend transition predicted a significantly lower intercept ($\beta = -.134$, $p < .05$) but higher growth rate ($\beta = .223$, $p < .01$) for the SAR construct. Thus, at the end of Tools implementation in the pre-kindergarten year, children whose teachers reported higher pretend transition implementation exhibited lower researcher-reported self-regulation, but those children then caught up with their peers over time through a faster growth rate. No other associations were observed.

Table 53: Beta coefficients for pretend transition on the three outcome measures

	Act → Intercept	Act → Slope
Executive function	-.054	-.027
CFBRS	.020	-.021
SAR	-.134*	.223**

Attention focus

Unlike the other activities in this section, attention focus does not refer to a discrete activity but rather to a set of chants, songs, and body movements. These activities aim to re-focus children's attention during a lesson or to gain their attention before a lesson. For example, the "Do what I do" game (Leong & Bodrova, 2011, p. 15) requires children to imitate a teacher's hand movements and speech patterns. The activity can increase in rigor as children have to match a teacher's pattern that the teacher is no longer acting out, which "requires even more working memory and inhibitory control" (Leong & Bodrova, 2011, p. 15).

In addition to this game, attention focus also refers to any other chants, songs, and body movements that teachers elect to use during the year. Neither the manual nor the PRI dataset distinguishes among all attention focus activities, so the activity had to be included in the analysis as a single predictor variable (i.e., instead of one variable for each of the songs, chants, and body movement activities).

Table 54 below indicates that the attention focus games predicted a significantly lower SAR intercept ($\beta = -.114$, $p < .05$) but a marginally higher growth rate ($\beta = .114$, $p = .08$). Thus, at the end of Tools implementation in the pre-kindergarten year, children whose teachers reported higher attention focus implementation exhibited lower researcher-reported self-

regulation, but those children then caught up with their peers through a faster growth rate. No other associations were observed.

Table 54: Beta coefficients for the attention focus activity on the three outcome measures

	Act → Intercept	Act → Slope
Executive function	-.034	-.041
CFBRS	-.089	.025
SAR	-.114*	.114 ⁺

Attention-focusing activity summary

Of the ten attention-focusing activities, three exhibited no significant associations with any self-regulation outcomes. By contrast, three of the four freeze game activities predicted lower intercepts (for freeze game and partner freeze) and slopes (for freeze on number) on at least one self-regulation outcome. The remaining four activities exhibited significant associations in one direction with a self-regulation intercept counterbalanced by a significant association in the other direction on the slope. Thus, none of the activities had an exclusively positive association with any self-regulation outcome. The next section presents the results for each Tools math activity, which brings the activity analyses section to a close.

12.4.3 Math activities

The final section below shares results for each of the twelve math activities and their individual associations with children's self-regulation. As with the previous models, all math activity models exhibited acceptable fit (RMSEA and SRMR under .08, CFI above .90). Moreover, none of the math activities exhibited collinearity issues, so all twelve are analyzed in the sections below.

Puzzles, manipulatives and blocks

In this activity, small groups of children learn to use puzzles, manipulatives and blocks alongside the teacher. Once the students have mastered independent use of the math learning materials, children can engage with the materials during free choice time without teacher intervention. Thus, this activity is a scaffold that eventually enables children to autonomously engage with the math learning materials (Leong & Bodrova, 2011, p. 413).

The puzzles, manipulatives, and blocks activity did not significantly or marginally predict any self-regulation intercepts or growth rates (see table 55 below).

Table 55: Beta coefficients for the puzzles, manipulatives, and blocks activity on the three outcome measures

	Act → Intercept	Act → Slope
Executive function	.036	-.066
CFBRS	-.021	.078
SAR	.041	-.141

Remember and replicate

Remember and replicate requires children to observe the teacher creating some geometric form with play dough, remember the form as the teacher hides it from view, and then replicate the form using their own play dough. The activity thus taxes children's geometric reasoning as well as their working memory, which illuminates the dual focus on academic and self-regulatory outcomes (Leong & Bodrova, 2011, p. 420). Remember and replicate did not significantly or marginally predict any self-regulation intercepts or growth rates (see table 56 below).

Table 56: Beta coefficients for remember and replicate on the three outcome measures

	Act → Intercept	Act → Slope
Executive function	-.086	.054
CFBRS	-.021	.027
SAR	-.042	.020

I have who has colors, shapes, and numbers

This set of three related activities requires children in small groups to use two-sided cards to develop fluency with shapes, numbers, and colors. An example of the shape game could proceed as follows: The teacher starts by announcing that his or her card has a square on it; next, a child in the small group announces that his or her card also has a square. That child then flips over the card and announces the shape on the card's other side (e.g., a triangle). After that, a child with a triangle card would announce that he or she has the triangle before

identifying the shape on the card's other side. The cards are designed such that each card set has exactly the right number of matches, and the game ends when all cards have been used (Leong & Bodrova, 2011, p. 430).

Thus, this activity targets both concept development (i.e., learning numbers, shapes, and colors) as well as self-regulation (i.e., inhibiting the impulse to call out before one's turn). Table 57 below indicates the intercept and growth rate estimates for each of the three 'I have who has' activities.

Table 57: Beta coefficients for I have who has colors, numbers, and shapes on the three outcome measures

	Colors		Numbers		Shapes	
	Intercept	Slope	Intercept	Slope	Intercept	Slope
Executive function	-.110 ⁺	.003	-.112 ⁺	-.012	-.103 ⁺	.045
CFBRS	-.108	.120	-.091	.114	-.040	.058
SAR	-.023	.021	-.006	-.045	.030	-.012

Table 57 indicates marginally significant and negative executive function intercepts for colors ($\beta = -.110$, $p = .064$), numbers ($\beta = -.112$, $p = .071$) and shapes ($\beta = -.103$, $p = .092$). Thus, at the end of Tools implementation in the pre-kindergarten year, children whose teachers reported higher 'I have who has' implementation exhibited lower executive function. No other significant associations were observed.

Making collections

This activity closely resembles buddy reading in that it involves a pair of children who switch roles mid-way through the activity. The making collections activity designates one child as the counter and the other child as the checker. The counter observes a card with a set of objects depicted on it (e.g., seven helicopters), places the appropriate number of counters on the card, and then allows the checker to verify their accuracy. After the children complete the set of cards, the two children switch roles and repeat (Leong & Bodrova, 2011, p. 441). Making collections did not significantly or marginally predict any self-regulation intercepts or growth rates (see table 58 below).

Table 58: Beta coefficients for making collections on the three outcome measures

	Act → Intercept	Act → Slope
Executive function	-.080	.012
CFBRS	-.044	.004
SAR	-.004	-.097

Math memory

Math memory requires children to observe a set of objects, recall the objects after a teacher hides them underneath a cloth, and then identify the emergent differences in the object set after the teacher has manipulated the objects under the cloth. Thus, this activity taxes children's spatial and geometric awareness as well as their working memory skills (Leong & Bodrova, 2011, p. 452). Math memory did not significantly or marginally predict any self-regulation intercepts or growth rates (see table 59 below).

Table 59: Beta coefficients for math memory on the three outcome measures

	Act → Intercept	Act → Slope
Executive function	-.005	-.009
CFBRS	.047	-.014
SAR	-.032	-.034

Number line hopscotch

This activity requires children to jump in the correct numbered sequence across different carpet squares (i.e., from square numbered one to square two to square three and so on). The teacher moves the squares after each round (Leong & Bodrova, 2011, p. 463). Thus, the activity taxes children's numerical awareness (i.e., knowing their numbers) as well as their cognitive flexibility and inhibitory control (i.e., to switch their jumping patterns according to the new carpet square arrangements).

Table 60 indicates that number line hopscotch predicted a significantly lower executive function ($\beta = -.113, p < .05$) and SAR ($\beta = -.115, p < .05$) intercept as well as a marginally

lower ($\beta = -.109$, $p = .097$) CFBRs intercept. Thus, at the end of Tools implementation in the pre-kindergarten year, children whose teachers reported higher number line hopscotch implementation exhibited lower executive function as well as teacher- and researcher-reported self-regulation. No associations with the growth rates were observed.

Table 60: Beta coefficients for number line hopscotch on the three outcome measures

	Act $\rightarrow$ Intercept	Act $\rightarrow$ Slope
Executive function	-.113*	-.077
CFBRs	-.109 ⁺	.054
SAR	-.115*	-.034

Numerals game

Numerals game also closely resembles buddy reading and making collections. The only difference between making collections and numerals game is that the latter involves a card with a number on it, whereas the former includes a card with a set of objects depicted on it. In both activities, the counter places the appropriate number of counters on the card, and then allows the checker to verify their accuracy. After the children complete the set of cards, then the two children switch roles and repeat (Leong & Bodrova, 2011, p. 470).

Table 61 below indicates that numerals game predicted a marginally lower executive function intercept ($\beta = -.115$, $p = .092$) and SAR growth rate ($\beta = -.162$, $p = .056$). Thus, at the end of Tools implementation in the pre-kindergarten year, children whose teachers reported higher numerals game implementation exhibited marginally lower executive function and also lower researcher-reported self-regulation growth through first grade. No other associations were observed.

Table 61: Beta coefficients for numerals game on the three outcome measures

	Act $\rightarrow$ Intercept	Act $\rightarrow$ Slope
Executive function	-.115 ⁺	-.044
CFBRs	-.087	.041
SAR	-.009	-.162 ⁺

Venger drawing

Venger drawing requires children to think of shape, draw the shape, and then imagine other shapes or objects into which that shape could be transformed. For example, children could draw a triangle, think of a bridge, and then begin to add shapes to their original triangle until they have created a bridge comprised of geometric shapes. This activity can also be completed in groups, which taxes children's behavioral control to take turns on determining the shape's evolution (Leong & Bodrova, 2011, p. 478).

Table 62 indicates that venger drawing predicted a significantly lower ($\beta = -.127$, $p < .05$) executive function intercept and a marginally lower ($\beta = -.119$, $p = .062$) CFBRS intercept. Thus, at the end of Tools implementation in the pre-kindergarten year, children whose teachers reported higher venger drawing implementation exhibited significantly lower executive function and marginally lower teacher-reported self-regulation. No associations with the growth rates were observed.

Table 62: Beta coefficients for venger drawing on the three outcome measures

	Act $\rightarrow$ Intercept	Act $\rightarrow$ Slope
Executive function	-.127*	-.024
CFBRS	-.119 ⁺	.040
SAR	-.092	-.019

Attribute game

Attribute game requires children to sort objects according to different attributes (e.g., color, size, shape, number of sides). Like many other Tools activities, children initially execute the attribute game in large groups, then smaller groups, then pairs, and, ultimately, individually. Thus, the activity taxes children's geometric awareness as well as their working memory, ability to take turns in groups, and, ultimately, self-regulated learning (Leong & Bodrova, 2011, p. 487). Attribute game did not significantly or marginally predict any self-regulation intercepts or growth rates (see table 63 below).

Table 63: Beta coefficients for attribute game on the three outcome measures

	Act → Intercept	Act → Slope
Executive function	-.057	-.027
CFBRS	-.077	.038
SAR	-.044	-.055

Patterns with manipulatives

This activity requires partner pairs to arrange manipulatives according to a pattern dictated by a colored strip. Each color on the pattern strip is associated, arbitrarily, with a certain manipulative (e.g., the color blue is represented by a small cube manipulative). One child replicates the pattern using the manipulatives, while the other child serves as the checker. After the children complete the strip, the switch roles and move onto more complex patterns throughout the year (Leong & Bodrova, 2011, p. 494).

Table 64 below indicates that the patterns with manipulatives activity predicted a marginally lower ($\beta = -.127$, $p = .055$) CFBRS intercept. Thus, at the end of Tools implementation in the pre-kindergarten year, children whose teachers reported higher patterns with manipulatives implementation exhibited marginally lower teacher-reported self-regulation. No other associations were observed.

Table 64: Beta coefficients for patterns with manipulatives on the three outcome measures

	Act → Intercept	Act → Slope
Executive function	-.102	-.065
CFBRS	-.127 ⁺	.148
SAR	-.003	-.110

Math activity summary

As with the make-believe play and attention-focusing activities, none of the twelve Tools activities exhibited exclusively positive associations with any of the three outcomes. Five activities had null associations across the three outcomes, whereas the remaining seven had negative associations with either the self-regulation intercept or slope. None of the twelve

activities exhibited any positive associations for either the intercept or the slope across the three outcome measures.

12.4.4 Brief sensitivity analysis with fidelity data

As described in section 11.7, Tools activity implementation data were collected in the form of both teacher-reported and researcher-observed measures. In the present study, the teacher-report measures served as the main measure for analysis because the PRI researchers only observed Tools classrooms for three days over the course of the year (Farran, Wilson, Meador, Norvell, & Nesbitt, 2015). Thus, the researcher-observed measure may be less representative of teachers' actual activity implementation throughout the year than the teacher-reported measure (Meador, 2015).

As in the activity block analyses in section 12.3, the researcher-observed data can be used for sensitivity analysis to check whether the results are consistent across both Tools implementation measures. Table 65 below shows the results for an overall fidelity composite score on the three self-regulation outcomes. That is, each teacher's fidelity scores across all Tools activities were averaged into a composite fidelity score for each teacher; this process enables the analysis of whether teachers with higher levels of Tools implementation fidelity are associated with improvements in children's self-regulation outcomes.

Table 65: Sensitivity analysis with the composite Tools teacher fidelity variable on the three outcome measures

	Executive function		CFRBS		SAR	
	Intercept	Slope	Intercept	Slope	Intercept	Slope
Composite fidelity	.05	-.16*	-.02	-.02	-.16***	.11+

As table 65 suggests, higher overall fidelity to Tools was not associated with higher self-regulation outcomes. In fact, higher fidelity across all Tools activities predicted a significantly lower ($\beta = -.16$, $p < .05$) executive function growth rate through first grade. Moreover, higher fidelity predicted significantly lower ($\beta = -.16$, $p < .001$) researcher-reported self-regulation skills at the end of the Tools implementation year (i.e., pre-

kindergarten) and marginally higher ($\beta = .11, p = .10$) researcher-reported self-regulation through the end of first grade.

In addition to the overall fidelity score analysis presented here, Appendix K includes the sensitivity analyses for each of the Tools activities. That is, every Tools activity represented in the fidelity data is tested for its association with the three outcome measures, which constitutes an exhaustive investigation of the Tools activities and children's self-regulation.

12.4.5 Summary of results for research question 4

Given that no previous study has examined the impact of specific Tools activities on children's self-regulation outcomes, the present study made no a priori hypotheses regarding their associations. Based on the sections above, it is clear that Tools activities mostly predicted either no change in children's self-regulation or had a negative association with the intercepts and growth rates. This trend was consistent across the make-believe play, attention-focusing, and math activities. Moreover, the sensitivity analysis using the composite fidelity variable mirrored the null and negative results from the main analysis using the teacher-reported implementation frequency data.

The only four activities that deviated from this trend include pretend transition, mouse trap, Mr. Wolf, and the generic attention focus activity. Those four activities exhibited significantly positive estimates for either the intercept or slope. However, those significantly positive estimates on one parameter were paired with significantly negative values on the other (i.e., if the intercept was positive, then the slope was negative and vice versa).

Thus, in line with Hypothesis three, children with significantly lower self-regulation at the beginning of pre-kindergarten exhibited steeper growth trajectories to catch up with other children or vice versa. No Tools activities exhibited an exclusively positive association with children's self-regulation. Table 66 shows only the sign for significant relationships observed in the activities above in order to visually summarize the findings.

Table 66: Observed effects for the play, math, and attention-focusing activities across the three outcome measures

	Executive function		SAR		CFBRS	
	I	S	I	S	I	S
Attention focus			-	+		
Attribute game						
Complete and continue						
Freeze game	-		-			
Freeze on number					-	
Graphics practice			-			
I have who has colors	-					
I have who has numbers	-					
I have who has shapes	-					
Making collections						
Math memory						
Make-believe play centers			-			
Make-believe play planning			-			
Make-believe play practice			-			
Mouse trap					+	-
Mr. Wolf activity					+	-
Number follow the leader						
Numeral games	-			-		
Number line hopscotch	-		-		-	
Partner freeze		-				
Patterns and manipulatives					-	
Pattern movement						
Pretend transition			-	+		
Puzzles with manipulatives						
Remember and replicate						
Two-step freeze						
Venger drawing	-				-	

Overall, these results do not provide evidence that any Tools activities should be applied across curricular contexts for the purpose of improving children's self-regulation. Two of the three make-believe activities, seven of the ten attention-focusing activities, and seven of

the twelve math activities predicted either lower intercepts or growth rates. Although no a priori hypotheses had been generated, it was expected that at least some of the activities would predict self-regulation improvement. For the make-believe play, attention-focusing, and math activities, this was not the case.

12.4.6 Covariate effects

By controlling for various covariates, the results reveal insights into the relationships among child characteristics and self-regulation development. Although those relationships are not included in this dissertation's research questions, the relationships merit a brief discussion. Future studies may be better able to select covariates based on the results here, or even pursue investigations of these covariates as the substantive focus of the research. The standardized coefficients of the covariates on the initial self-regulation status and growth rate for the executive function, CFBRS, and SAR constructs are presented in tables 67 through 69, respectively.

Covariates and the executive function construct

Table 67 indicates significantly lower executive function intercepts at the beginning of pre-kindergarten for English language learners (ELL), males, and students with special educational needs (SEN). By contrast, older children have significantly higher executive function intercepts. As for the slope parameters, ELL and SEN students grow faster than average to catch up over time, whereas older children's growth trajectories shallow to eventually mirror the trajectories of their younger classmates.

Table 67: Covariate effects on the executive function intercept and growth parameter estimates

	Intercept	Slope
English language learner (0 = No, 1 = Yes)	-.283 ^{***}	.342 ^{***}
Gender (0 = girl, 1 = boy)	-.125 ^{***}	.016
Special educational needs (0 = No, 1 = Yes)	-.204 ^{***}	.120 [*]
Age in months (centered)	.272 ^{***}	-.148 ^{***}

Covariates and the teacher-reported self-regulation (CFBRS) construct

Table 68 indicates significantly lower teacher-reported self-regulation intercepts for male and SEN students. By contrast, older children have significantly higher CFBRS intercepts. As for the slope parameters, ELL students grow faster than average through first grade. SEN students exhibit a marginally significant positive growth rate, whereas older children exhibit a marginally negative one. Both of those marginally significant effects are in the opposite direction from the intercept estimate; this association aligns with the expectation articulated in Hypothesis 3 about an inverse relationship between intercepts and slopes.

Table 68: Covariate effects for the CFBRS intercept and growth parameter estimates

	Intercept	Slope
English language learner (0 = No, 1 = Yes)	.040	.243***
Gender (0 = girl, 1 = boy)	-.229***	-.068
Special educational needs (0 = No, 1 = Yes)	-.219***	.088 ⁺
Age in months (centered)	.224***	-.074 ⁺

Covariates and the SAR construct

Table 69 indicates significantly lower researcher-reported self-regulation intercepts for male and SEN students. By contrast, ELL students and older children have significantly higher SAR intercepts. As for the slope parameters, boys grow faster than average through first grade to catch up with their female classmates. No other associations were observed.

Table 69: Covariate effects for the SAR intercept and growth parameter estimates

	Intercept	Slope
English-language learner (0 = No, 1 = Yes)	.137***	-.010
Gender (0 = girl, 1 = boy)	-.191***	.135**
Special educational needs (0 = No, 1 = Yes)	-.249***	-.009
Age in months (centered)	.171***	-.029

Covariate effects summary

The observed covariate effects aligned with common trends in the literature. Specifically, boys and SEN students mostly exhibited diminished self-regulation skills relative to girls and non-SEN students, respectively. By contrast, older children exhibited significantly more robust self-regulation skills than younger children across the three outcome measures.

Finally, ELL students exhibited inconsistent patterns; specifically, ELL students had lower executive function intercepts versus equal or higher scores on the two informant-report self-regulation assessments. This is likely because the task-based executive function measures require close attention to complex directions provided in English, whereas the subjective report measures were unrelated to children's English language proficiency.

12.5 Summary of results for study two

This chapter shared the results for study two, which covered four topics:

- Children's self-regulation development from pre-kindergarten through first grade
- The overall effect of Tools on children's self-regulation development vis-à-vis comparison curricula
- The associations between groupings of Tools activities (e.g., make-believe play, mathematics, attention-focusing games) and children's self-regulation growth
- The associations between individual Tools activities and children's self-regulation

The results from each section, respectively, were as follows: Firstly, self-regulation skills mostly increased, on average, for the 1140 children in the sample between pre-kindergarten autumn and first grade spring. Secondly, Tools students had significantly shallower executive function growth trajectories than those of comparison group students. Thirdly, the make-believe play, math, and attention-focusing activity blocks exhibited significantly negative associations with self-regulation development, whereas the other activity blocks exhibited no associations. Fourthly, several individual activities exhibited negative and significant associations with self-regulation development, while no activities predicted consistently higher self-regulation skills.

The implications of these results are described in the upcoming discussion chapter. The discussion chapter then transitions into the fourteenth and final chapter, which concludes the dissertation by synthesizing study one and two's findings and discussing the findings' implications for policy and practice.

CHAPTER 13: Discussion for study two

This chapter discusses the findings for study two, which examined Tools of the Mind's (Tools) overall effect on self-regulation as well as the associations between Tools activities and self-regulation. Whereas Chapter Twelve presented the results for study two, the current chapter contextualizes those results within the existing literature base.

In the sections below, the results for the four research questions of study two are discussed and situated within the extant research base. After that, study two's strengths and limitations are assessed. Finally, directions for future research are identified before transitioning to the final chapter, which describes these findings' implications for policy and practice and brings the thesis to a close.

13.1 Research question one: Children's self-regulatory development between pre-kindergarten and first grade

As described in the research hypotheses section (11.10.1), the phenomenon of children's self-regulatory developmental trajectories can be divided into at least three sub-questions:

1. On average, does children's self-regulation decline, grow, or stagnate across time?
2. How much variability in those developmental trajectories is observed within and across children?
3. What is the relationship between the initial self-regulation status and the self-regulation growth rate? That is, do children with low self-regulation have a higher growth rate? That is, do the children with initially low self-regulation, on average, catch up with their peers and vice versa?

Each of those questions is sequentially addressed in the sections below.

13.1.1 Positive growth in self-regulation on average

Hypothesis 1 (see section 11.10.1) predicted that children's self-regulation would, on average, increase between pre-kindergarten and first grade. The results from section 12.1 support

Hypothesis 1 (i.e., positive self-regulation growth) across six of the seven self-regulation measures for the 1140 children in the sample. That finding is consistent with previous studies that indicate positive self-regulation growth through the end of pre-kindergarten at age five (Kopp, 1982; Moilanen et al., 2009) as well as the end of first grade at age seven (Eisenberg, Valiente, Morris, Gershoff, & Shepard, 2003; Eisenberg et al., 2005).

Neuroscience research indicates (Blair, 2006; Diamond, 2006) that the pre-frontal cortex, which mediates self-regulatory capacity, is growing rapidly during the early childhood years. As the anterior cingulate cortex, which is within the pre-frontal cortex and most directly links to self-regulation (see section 5.1.2) develops in the early years, children's cognitive and behavioral regulation capacities develop in kind (Bell & Deater-Deckard, 2007).

In addition to biological changes, contextual influences such as the transition to school also promote self-regulation growth. Specifically, the comparatively structured nature of a classroom environment hones children's self-regulatory skills as they learn to remain focused and persist on academic tasks (Barnett et al., 2008; Duckworth & Carlson, 2013). Of course, children experiencing negative contextual influences such as poor schooling, poor parenting, or poverty would be likely to experience less self-regulatory growth. That is why the observed result of positive growth is simply the *average* trend.

In fact, one observed exception to the positive self-regulation trend is the teacher-reported Cooper-Farran Behavioral Rating Scale (CFBRS) construct, where growth leveled off by kindergarten and became negative between kindergarten and first grade. Because CFBRS scores derive from teacher reports, it is possible that kindergarten and first grade teachers had higher self-regulatory expectations for students than did the pre-kindergarten teachers. Although this hypothesis could not be systematically tested with the available data, existing research (Lin, Lawrence, & Gorrell, 2003; Rimm-Kaufman et al., 2000) indicates that some teachers, especially in early childhood (Charlesworth et al., 1993), have developmentally inappropriate expectations regarding children's self-regulation relative to natural and appropriate levels of development.

Thus, teachers' expectations may explain the observed decline in teacher-reported CFBRS, whereas the results from the other six self-regulation indicators support Hypothesis 1 of

positive growth across time. Once again, though, the positive growth trend is only an average trajectory, whereas the results discussed in section 13.1.2 below demonstrate the substantial variability across children around that average.

13.1.2 Variability of self-regulation trajectories among children

In line with Hypothesis 2 (see section 11.10.1), although the indicators mostly indicated average growth across time, six of the seven models also exhibited substantial inter-individual variability in both initial self-regulation status and growth. That is, different children had significantly different patterns of self-regulation change over time. The Dimensional Change Card Sort (DCCS) task emerged as the only exception, where children showed highly similar growth patterns on this task over time. Across the other six indicators, though, the significant overall variability suggests that many children's individual self-regulation trajectories diverged substantially from the sample average.

The observed inter-individual variability conforms with previous research (Bronson, 2000; Eisenberg et al., 2005) that suggests several possible trajectories of self-regulation development in both early childhood and adolescence. Irregular self-regulatory trajectories have been found in normally developing young children (Dennis, 2006). Given that children's developmental trajectories for many capacities, not just self-regulation, can vary dramatically (Skinner, Zimmer-gembeck, & Connell, 1998), the observed variability in self-regulation growth aligns with developmental expectations.

Additionally, children from impoverished families are more likely than their advantaged counterparts to exhibit substantial variability in self-regulation growth trajectories (Raver et al., 2013), which may reflect the relative instability in their daily lives (Bronson, 2000). Given that the 1140 children in this sample were assumed to be mostly high-poverty (see section 11.8), it is further unsurprising that the present results support Hypothesis 2 regarding significant variability across children.

13.1.3 Association between the initial self-regulation status and growth rate

Hypothesis 3 predicted a negative association between the initial self-regulation status and growth rate. That is, children with low initial self-regulation at the beginning of pre-

kindergarten were expected to exhibit steeper self-regulation growth (i.e., catch up) relative to those with high initial self-regulation levels. Conversely, children with high initial self-regulation at the beginning of pre-kindergarten were expected to exhibit shallower self-regulation growth through first grade relative to children with low initial self-regulation.

The results from section 12.1 only partially support Hypothesis 3. Four of the seven outcome measures indicate a negative association between the initial self-regulation status and growth rate. Specifically, both of the informant-report measures (i.e., the teacher-reported CFBRs and the researcher-reported SAR) indicated a negative association, as did the Peg Tapping and Heads-Toes-Knees-Shoulders tasks.

For the two informant-report measures, the negative association between initial self-regulation status and growth rate could be explained by the scale of the measures. That is, both informant-report measures (the SAR and CFBRs, respectively) contain five ordinal indicators each scaled from 1-5. For both measures, some children received high scores from the beginning of pre-kindergarten (i.e., the first testing administration). Thus, children with a high initial status had less room to grow than those who had earned a lower score at the beginning of pre-kindergarten. Thus, the negative association for the informant-report measures may be explainable due to the ceiling effect, whereby children's increasing age (the independent variable) leads to no change in self-regulation (the dependent variable) because the child has already reached the score ceiling at the study's outset.

The findings of a negative association between initial status and growth rate aligned with Baron & Malmberg (2014), who also found a negative association between initial self-regulation status and growth rate using nationally representative data from the Millennium Cohort Study, as well as Moilanen (2009), who found a negative association between initial self-regulation status and growth rate with a sample of 731 American children.

As for the present study's remaining three self-regulation measures, no association was observed between initial status and growth rate for Copy Design, Corsi Blocks, or DCCS. Therefore, none of the seven measures supported Skinner's (1998) launch model, which would predict that higher initial self-regulation status would 'launch' accelerated self-

regulation growth across time. Conversely, the results mostly supported Hypothesis 3, whereby children with lower initial self-regulation status ‘catch up’ to their peers across time.

13.2 Research question two: Tools’ effect on self-regulation vis-à-vis comparator curricula

The second research question shares the same analytic aim as the meta-analysis in study one, which sought to evaluate Tools’ overall effectiveness vis-à-vis comparator programs.

Whereas the meta-analysis used data from all existing Tools studies, this second research question analyzed Tools’ effectiveness only using data from the Peabody Research Institute’s (PRI) large-scale cluster-randomized Tools trial (Farran & Wilson, 2014).

The present study goes beyond the Farran & Wilson (2014) study in three ways:

1. This study used latent variable modeling to account for measurement error in the self-regulation constructs.
2. The latent growth models in this study enabled assessment of inter-individual variability in self-regulation growth, whereas the PRI study’s regression approach precluded analysis of developmental heterogeneity across children.
3. The present study assessed differences across the two chronologically sequential cohorts of Tools children (see section 11.2), which was not reported in the previous PRI study using these data (Farran & Wilson, 2014).

Although the present analysis improved upon Farran & Wilson (2014) in the aforementioned three areas, the results in this study largely mirrored those of the PRI study. Specifically, children randomly assigned to Tools classrooms exhibited no significant self-regulation differences from comparison group children at the end of the Tools implementation year (i.e., pre-kindergarten), which did not support Hypothesis 4 of higher self-regulation skills for Tools children.

Moreover, this study found a significantly lower executive function growth rate through first grade for children randomly assigned to Tools classrooms, which further accords with

Farran & Wilson's (2014) findings of lower composite self-regulation scores for Tools children at the end of first grade.

In the present study, Hypothesis 4 had predicted significantly higher self-regulation skills for Tools children based on the supposition that Farran & Wilson's (2014) counterintuitive results may have been due to measurement error in the self-regulation indicators. However, the present study's results contravened Hypothesis 4 by showing significantly *lower* executive function for Tools children, despite the fact that I used latent constructs to account for measurement error and also analyzed the full sample (i.e., both cohorts) of children.

Thus, the findings for this research question both align with and extend upon the results observed in the meta-analysis from study one. That is, the null findings at the end of Tools pre-kindergarten implementation mirror the null findings for self-regulation observed in the meta-analysis. Moreover, the negative effects of Tools on children's executive function growth through first grade go beyond the meta-analytic findings to suggest that Tools may actually impair self-regulation development over time.

Once again, these findings can be considered causal because schools in the Farran & Wilson (2014) study were randomly assigned to either Tools or comparison curricula. Given this experimental research design, the findings of this study suggest that Tools may not deliver on its stated aim of improved self-regulation skills and may in fact compromise the very quality it seeks to cultivate.

13.3 Research question three: Subject-specific Tools activity blocks and self-regulation

Even though the results indicate that Tools, on the whole, does not improve children's self-regulation skills more than other curricula, it remains important to assess whether certain elements of the curriculum do promote such skills. In so doing, effective Tools activities can be identified and emphasized more strongly by the curricular developers and Tools teachers, whereas ineffective Tools activities could be deemphasized or abandoned. Theoretically, such granular refinements could improve the overall program's effectiveness.

Thus, research question three analyzed whether subject-specific Tools activity blocks²¹ (e.g., literacy, math, make-believe play) are associated with self-regulation.

Whereas previous studies (Baron & Malmberg, 2014; Farran & Wilson, 2014; Moilanen et al., 2009) have already addressed the aims of research questions one and two above (i.e., self-regulation developmental trajectories and Tools' overall effectiveness), research question three involves a novel analysis heretofore unseen in the literature. That is, to the best of my knowledge, no study has analyzed specific Tools activities. Consequently, no previous studies bear directly on the findings presented here.

Nonetheless, the results of this dissertation's activity block analysis can still be contextualized within literature regarding the linkages between various instructional activities and children's self-regulation. It is important to note that, unlike the research question two results, which can be considered causal due to the randomization of students to curricula, the results for this research question are correlational. This is because the Tools students ($n = 646$) were not randomly assigned to various blocks of Tools activities. Nonetheless, the results for each subject-specific block are reviewed below against existing literature from other non-Tools curricula and educational settings.

Make-believe play activity block

According to the Tools curricular developers, "the best way for children to practice self-regulatory behaviors is to engage in mature make-believe play" (Bodrova & Leong, 2015a). The present dissertation is the first to actually test this claim using Tools data. As the results in section 12.3.1 indicated, Tools students ($n = 646$) who experienced higher levels of Tools make-believe play implementation did not exhibit higher levels of self-regulation. In fact, higher make-believe play block implementation was associated with marginally lower executive function growth through first grade and significantly lower researcher-reported self-regulation skills at the end of pre-kindergarten (i.e., the Tools implementation year).

²¹ Once again, I used the Tools manual (Leong & Bodrova, 2011) to determine appropriate groupings of Tools activities. The manual includes clearly delineated sections for different subjects and enumerates a set of activities that fit into each subject designation.

Although the findings regarding make-believe play do not align with the 'Tools developers' expectations (Bodrova & Leong, 2007) or my (section 11.10) hypotheses, the findings do align with recent research (Carlson et al., 2014; Lillard et al., 2013) that indicates mixed evidence regarding the association between pretend play and self-regulation.

Once again, the Lillard et al. (2013) systematic review described in section 5.2.3 found inconsistent associations between pretend play and self-regulation. While some of the studies reviewed by Lillard et al. (2013) indicated positive associations between play and self-regulation, many others indicated null or even negative associations. Lillard et al. (2013) concluded that evidence of self-regulation promotion through pretend play is “sparse at best” (p. 23), despite claims from others in the field (Bodrova & Leong, 2013; Weisberg et al., 2013a) that Lillard et al. (2013) overstated their claims through selective criticisms of the included play literature.

In the most recent study I could find pertaining to play and self-regulation, Faja et al. (2016) found that children's early executive function skills do predict their levels of play (measured through an unstandardized play assessment), but, conversely, children's early play did not predict subsequent executive function. The authors concluded that the relationship between executive function and play is “specific” (p. 6), where executive function promotes play but not the other way around. If this is indeed the case, then the 'Tools developers' claim would be exactly backward – self-regulation can improve the quality of children's play, but play may not improve their self-regulation.

Another possible explanation for the present study's findings lies with Vygotsky's (1933b) notions of mature, or structured, play. According to one of Vygotsky's followers, Elkonin (1975), make-believe play must be “fully developed” and in a “mature” (p. 17) stage in order for it to effectively cultivate children's self-regulation. Similarly, the curricular developers argue (Bodrova, Leong, & Akhutina, 2011) that due to changes in modern society, children engage in pretend play less than ever before; consequently, children “may engage in play that never reaches its mature level” (p. 21). Consequently, it could be argued that children's self-regulation should not be expected to improve without those high levels of mature play.

This explanation may indeed be plausible. On the other hand, though, the promotion of mature and structured play is a central goal of Tools (Bodrova & Leong, 2007). If play did not reach its mature level with a relatively representative group of 80 teachers in 59 schools in the United States, then it is difficult to imagine other mainstream contexts in which Tools' play would reach necessary levels of maturity.

Moreover, it may be that expecting structured play from preschool children, who are expected to “constantly monitor each other to make sure that everyone is following the rules” (Bodrova & Leong, 2003, p. 13), is unrealistic. Even Vygotsky, in his 1933 lecture on play, asserted “Whenever there is an imaginary situation in play, there are rules – *not rules that are formulated in advance*” (1933b, p. 10). Thus, the Tools' developers decision to integrate highly structured make-believe play planning processes before the play scenario may not even accord with Vygotsky's own prescriptions for effective play.

Relatedly, it is critical to remember that this dissertation tested Tools' make-believe play approach specifically as opposed to pretend play more generally. Although pretend play did not predict higher self-regulation in this study, that result cannot be generalized beyond the Tools context, which, again, has a highly structured and specific approach to make-believe play in early childhood. Many studies have demonstrated (Berk & Meyers, 2013; Elias & Berk, 2002; Rosen, 1974; Weisberg et al., 2013b) benefits of play for a range of developmental outcomes, whereas the negative results observed in this study pertain only to the Tools make-believe play context.

Math activity block

Abundant research (Bull & Scerif, 2010; Fuhs et al., 2015; Nunes et al., 2007) links children's executive function skills with higher math achievement levels. The question addressed here, however, is whether math instruction can reciprocally promote children's executive function. In the case of the present study, the results from section 12.3.3 indicate that more math block implementation predicts marginally lower executive function at the end of pre-kindergarten and has no association with any other outcome at any time point for Tools students ($n = 646$).

Some might argue that despite evidence that self-regulation skills improve math skills, it is unreasonable to expect that math instruction will improve self-regulation. In fact, Clements et al. (2014) tested this exact hypothesis in a randomized controlled trial using a math intervention called Building Blocks (Sarama & Clements, 2003). Not only did the authors test the association between the Building Blocks math curriculum and executive function, but they also tested whether augmenting the Building Blocks math curriculum with the Tools program would further enhance children's executive function.

In a sample of over 1000 American children, Clements et al. (2014) found that the Building Blocks math curriculum significantly improved children's executive function; however, in another group that received both the math curriculum combined with elements of Tools, the children exhibited no significant executive function gains above the 'business-as-usual' comparison group (Clements et al., 2014, p. 45). Thus, not only did Clements et al. (2014) demonstrate that math curricula can improve self-regulation skills, but they also demonstrated that infusing Tools elements into the Building Blocks math program actually suppressed the math program's positive effect on executive function.

Of course, it remains possible that simultaneously implementing two curricula was simply too burdensome for teachers, and that this explains the null effects for the Building Blocks – Tools combined intervention. Moreover, because Clements et al. (2014) did not test a Tools-only group against Building Blocks, the relative efficacy of the two curricula could not be directly analyzed. Nonetheless, Clements et al. (2014) did demonstrate that the Building Blocks math curriculum could improve executive function, which was corroborated in another American study (Weiland & Yoshikawa, 2013) and elaborated upon more recently in Clements, Sarama, & Germeroth (2016).

Notably, the Tools developers identify (Leong & Bodrova, 2011) math as an area of weakness for the curriculum, explaining that they specialize more in literacy instruction than in math. However, Tools does seek to embed self-regulation training into all activities, including math, which we might expect to have some positive effect on children's self-regulation outcomes as was observed in the study one meta-analysis. Moreover, since the Building Blocks curriculum (Sarama & Clements, 2003) has been shown to promote

children's self-regulation skills, it seems that 'Tools' math block could be reasonably expected to improve self-regulation given that this is 'Tools' explicit goal.

Literacy activity block

Similar to the math section above, extensive research (Blair & Razza, 2007; Fuhs, Nesbitt, Farran, & Dong, 2014; Nix et al., 2013; Stechuk, 2008b) indicates that self-regulation skills promote literacy learning. Again, though, the present question is whether literacy activities can reciprocally promote children's self-regulation. The results from section 12.3.2 of this dissertation indicate no significant associations between the Tools literacy block and any of the three outcome measures for Tools students ($n = 646$).

As with the math section above, we might ask whether literacy activities should be reasonably expected to improve children's self-regulation. In a recent meta-analysis of 67 studies regarding school-based executive function interventions and academic achievement, Jacob & Parkinson (2015) found that reading achievement exhibited a significantly lower ($\beta = -.12, p < .01$) association with executive function than did math (p. 17).

Given the weaker relationship between literacy and executive function, it may be unlikely that a literacy intervention would improve self-regulation. Indeed, in my review of the literature, the vast majority of literacy studies (Fuhs et al., 2015; McDonald et al., 2011; Skibbe, Phillips, Day, Brophy-Herb, & Connor, 2012; Stechuk, 2008) indicated no evidence that literacy learning directly improved children's self-regulation.

In fact, I found only one study (Weiland & Yoshikawa, 2013) that linked a literacy curriculum to executive function growth. Because the intervention group children in this study experienced both a literacy curriculum, Opening the World of Learning (OWL; Schickedanz & Dickinson, 2005), as well as the Building Blocks math curriculum referred to above, it is impossible to know whether the literacy curriculum was in fact responsible for the executive function gains among intervention group students.

Thus, although the authors link the OWL literacy curriculum to executive function gains (Weiland & Yoshikawa, 2013, p. 2126), the study design precluded empirical assessment of this hypothesis. Thus, to the best of my knowledge, it is difficult to conclude that any

literacy curricula exhibit evidence of effective self-regulation promotion. Consequently, the fact that the Tools literacy block does not exhibit any association with children's self-regulation mirrors the evidence from the existing literature on literacy curricula.

Science activity block

In contrast to math and literacy, which both have rich self-regulation-related literature bases, very few studies specifically pertaining to science education and self-regulation exist, especially in the early childhood literature. This is probably because science is not often explicitly taught during early childhood (Tough, 2012) and also because literacy and math have been studied more thoroughly than science across the age spectrum.

Those limitations aside, it remains important to discuss the findings from study two regarding science and self-regulation. Specifically, the study two results indicated no significant associations between science and Tools students' ($n = 646$) self-regulation skills. Importantly, though, this science activity block was not truly a 'block' (i.e., an average of multiple activities) because Tools has only one science activity. If Tools expanded its science instruction to include more activities, then perhaps the association between Tools science instruction and children's self-regulation would strengthen.

Again, though, the literature in this area is quite limited, so it remains difficult to make any strong claims about whether science instruction can cultivate self-regulation. One study entitled "Promoting self-regulation in science education" (Schraw et al., 2006) pertains to middle school science and presents a theoretical framework for self-regulation promotion; the study lacks any empirical data to demonstrate that science education does, in fact, hone self-regulation skills.

Although other studies indicate that science teaching can improve higher-order thinking (Weaver, 1998), creativity (Baron & Chen, 2012), and problem-solving (Huffman, 1997), I could not find research that empirically linked science teaching with self-regulation growth. More research is thus necessary in this area to determine how self-regulation promotion could be effectively embedded within science instruction.

Introduction activities and attention-focusing activities

The introduction and attention-focusing activity blocks included activities that are common in early childhood classrooms (e.g., freeze dance, name games) but made distinctive and structured within the Tools framework (e.g., holding up Tools body position cards for children to imitate when the music freezes, playing a Tools name game that designates the order in which students introduce themselves). As with the other activity blocks above, neither the introduction nor attention-focusing activity blocks predicted improved self-regulation among the Tools students ($n = 646$).

Specifically, introduction activities had no association with any of the three self-regulation outcome measures, whereas attention-focusing activities predicted a marginally lower researcher-reported self-regulation intercept (see sections 12.3.5 and 12.3.6, respectively). Because the form of these activity blocks is highly specific to Tools, no existing research bears directly on these findings.

Nevertheless, the findings of several existing studies do have indirect implications for the present findings. Firstly, previous studies (Diamond, 2012; Tominey & McClelland, 2011) have found that games involving inhibitory control can improve children's self-regulation. For example, the game Simon Says requires children to imitate a leader's body movements only when the leader first exclaims "Simon Says." Thus, children must inhibit the impulse to imitate the leader's movements on every turn and instead only do so when the relevant rule (i.e., "Simon Says") is applied. The original study on Simon Says (Strommen, 1973) found improved impulse control among children who played the game, though that study involved no randomization process to rigorously test the game's effectiveness.

Similar to Simon Says, each Tools activity in the attention-focusing and introduction activities section is designed to tax executive function. Given that each of the Tools activities contained similar elements to tax children's executive function, the question arises: Why didn't the activities effectively improve children's self-regulation skills? One possibility is that the Tools activities were actually *excessively structured*. That is, while Tools aims to progress from external regulation by teachers to children's autonomous self-regulation, the

Tools activities may actually represent highly structured forces that externally regulate children's behavior, thus undermining children's ability to regulate themselves.

For example, many traditional early childhood classrooms feature a freeze dance activity, where children dance to music and then freeze when the music stops. Whereas the traditional freeze dance game involves just one rule (i.e., freeze when the music stops), the Tools freeze activity involves much more structure. Specifically, the Tools 'two-step freeze' game includes two body pose cards that are color-coordinated. As music plays, children briefly view both body pose cards. The cards are then removed by the teacher, who holds up a color card, which corresponds to only one of the two body position cards. When the music stops, children must strike the appropriate body pose that corresponds to the color card previously held aloft by the teacher and inhibit the impulse to strike the other pose (Leong & Bodrova, 2011, p. 35).

Of course, the justification for the multi-step freeze game in Tools is that those steps directly tax children's self-regulation skills. However, the converse may be true; that is, in seeking to impose 'attention-focusing activity' structures onto each part of the day, the Tools program may actually be threatening the very capacity it seeks to promote.

One recent study (Barker et al., 2014) specifically investigated whether more-structured or less-structured activities in the home predicted higher executive function. Using correlational analyses, the authors found that more-structured activities predicted significantly lower executive function whereas less-structured activities predicted significantly higher executive function among a sample of 70 children ages six and seven. Ironically, the authors conclude that "consistent with Vygotskian developmental theory and programs that build on that theory, such as Tools of the Mind, less-structured time may uniquely support the development of self-directed control" (Barker et al., 2014, p. 10).

Despite Barker et al.'s (2014) claims about Tools, the regimented nature of activities such as freeze dance, make-believe play, buddy reading, and others belie the notion that Tools is relatively unstructured. While several preschool classrooms feature activities such as make-believe play, Tools is unique in its attempt to formalize the play scenario to require children to write and adhere to plans about their play. And, again, while many preschool classrooms

feature freeze dance games, Tools is unique in its requirement that children perform a specific dance move depicted on a specific color-coordinated card when the music stops. This list could go on for the other 59 Tools activities, but the underlying logic should hopefully be clear: highly structured activities may retain the regulatory function within the activities themselves rather than enabling children to internalize that regulatory function.

Again, this hypothesis regarding structured versus unstructured activities was not rigorously tested in the present study, so the legitimacy of the claim cannot be evaluated. Thus, it is entirely possible that the structured nature of the activities bears no relation to the results. Instead, some might expect that factors such as low fidelity could explain the disappointing results for the Tools classrooms. However, the fidelity analyses in this study (see section 12.4.4) as well as the previous PRI study (Farran & Wilson, 2014) suggest that low fidelity was not the source of the problem; again, the section 12.4.4 results indicate that high fidelity to Tools predicted significantly lower self-regulation.

Activity block summary

In sum, none of the activity blocks exhibited exclusively positive associations with children's self-regulation. Instead, the activity block analyses mostly indicate null or negative associations with children's self-regulation. Again, the explanation behind these results could be that the Tools activities are actually excessively structured, which served to externally regulate children's behavior instead of promoting children's autonomous regulation over themselves. Alternatively, it remains possible that the observed negative results derive from a range of common issues such as poor professional development training, insufficient funding, or other factors, which could all be explored in future research.

13.4 Research question four: Specific Tools activities and self-regulation development

Whereas the results section for research question four was the longest in Chapter Twelve, the present discussion section will be the shortest. This is because no previous study has analyzed individual Tools activities' association with self-regulation outcomes, or any outcomes for that matter. Whereas the Tools activity block analyses from the previous

section could be compared against similar subject curricula in other settings (e.g., Building Blocks for math or OWL for literacy), each discrete Tools activity is specific to (and corporately trademarked by) the Tools program. Thus, the findings cannot be contextualized within previous literature that does not exist.

That said, the results can be briefly summarized as follows: All but one of the individual Tools activities had either null or negative associations with the outcome measures (see Appendix K for the full results across all activities). The weather graphing activity, where children and teachers chart the daily weather patterns, was the only activity that showed exclusively positive associations with self-regulation, even though a theoretical explanation for that effect remains unclear (i.e., it could simply be a Type I error).

Some activities exhibited a positive association with the initial self-regulation status at the end of pre-kindergarten, which was then counteracted by a negative growth rate through the end of first grade (or vice versa). Thus, even if the activity predicted higher initial self-regulation, that gain was nearly always counterbalanced by slower self-regulation growth relative to their peers through the end of first grade. Overall, then, it is not clear whether any Tools activities could be replicated across other curricular contexts to improve children's self-regulation.

13.5 Brief summary of the study two findings

The findings from study two suggest that the Tools program as well as the individual activities that comprise it do not uniquely promote self-regulation. Whereas the Tools activity analyses are correlational, the analysis of the program as a whole (see section 12.2) employed a cluster-randomized design to produce causal findings. Those causal findings demonstrate that children in Tools classrooms exhibited significantly lower executive function growth through first grade, whereas Tools did not significantly affect children's teacher- and researcher-reported self-regulation at any time point.

That finding was further bolstered by the activity analyses, which mostly indicated null or negative associations between Tools activities and self-regulation. Those negative associations emerged both for blocks of Tools activities as well as the individual activities on

their own. Only one Tools activity (i.e., weather graphing) consistently predicted higher self-regulation for Tools students.

Although study two's findings mostly represent novel contributions to the literature, this research is not without its limitations. After briefly presenting three strengths of this research in section 13.6 below, section 13.7 details this study's limitations, which could provide useful guidance for future Tools research.

13.6 Strengths

This study has three noteworthy strengths. The first is the quality of the data sample; the second is the sophistication of the statistical approach; and the third is the exhaustive analytic approach. Each strength is briefly discussed in turn below.

13.6.1 Quality of the data sample

First, the data for study two derived from a large-scale project funded through a grant from the U.S. Department of Education (Farran & Wilson, 2014). Thus, the dataset is far more expansive than an individual researcher could independently collect. In addition to its size, the dataset's quality has been noted by self-regulation scholars such as Diamond (2012), who called the PRI study "particularly noteworthy because of its impeccable research design and the meticulous way in which it is being conducted" (p. 338).

I would not argue that the PRI research design is impeccable, as will be discussed in the limitations section, but the breadth of the data collection was truly noteworthy. Specifically, the combination of a five-part executive function test battery alongside researcher- and teacher-reported self-regulation over a five-year longitudinal study is both rare and valuable. Such rich data enabled deep analysis of the self-regulation construct, which explains why I pursued this dataset from the outset.

13.6.2 Sophistication of the statistical approach

A second strength of this study is the sophistication of the statistical approach. By modeling the self-regulation indicators as latent constructs, the measurement error problem that often

plagues self-regulation research (McClelland & Cameron, 2012; Miyake et al., 2000) was effectively addressed. Moreover, by testing the latent self-regulation constructs within latent growth models, the analyses went beyond traditional approaches (e.g., repeated-measures ANOVA) by modeling intra- and inter-individual differences in self-regulation development as substantively interesting rather than as error variance.

Additionally, the nested nature of these data warranted the use of multilevel structural equation models, of which latent growth models are a part, which corrected for the autocorrelation for children in the same classrooms. Overall, this rigorous methodological approach yielded maximally accurate estimates of Tools curricular effects on children's self-regulation skills.

13.6.3 Exhaustive analytic approach

The third strength of this study is the exhaustiveness of the analytic approach. Specifically, the PRI dataset afforded access to information on both teacher-reported activity implementation as well as researcher-reported activity implementation. The analyses incorporated both measures, which involved the execution of over 250 distinct models (i.e., the 45 teacher-reported activity items as well as the 40 researcher-observed items tested against all three outcome measures over time).

Moreover, every self-regulation outcome measure was tested first through an exploratory factor analysis, a confirmatory factor analysis, and a longitudinal measurement invariance model before being entered into the multilevel structural equation modeling analyses. Consequently, the analyses represent a comprehensive investigation of Tools' curricular effects on self-regulation, which ranks as the primary aim of this dissertation.

13.7 Limitations

Although the analysis exhibits multiple strengths, the present study has limitations of both the data sample and the data analyses. The seven limitations will be divided below according to 1) data sample limitations and 2) analytic limitations.

13.7.1 Data sample limitations

Potential bias in the subjective self-regulation measures

In study two, teachers and researchers rated children's self-regulation using the CFBRs and SAR scales, respectively. Both teachers and researchers were aware of the children's condition assignment (i.e., Tools or comparison condition) when filling out the scales during the pre-kindergarten year. This situation is unavoidable for teachers because they know which curriculum they are using, whereas the PRI research assessors could have been blinded from the experimental condition of the students.

Thus, the lack of blinding emerges as a source of bias. Raters who are evaluating Tools students may be expecting Tools children to have improved self-regulation because of their engagement with a self-regulation oriented curriculum. By contrast, raters may have lower self-regulatory expectations for comparison group children. Thus, the subjective expectations of teachers and researchers could have possibly biased their ratings, which could further bias the model results.

Evidence of this phenomenon was observed in the results. Although several negative associations with the objective executive function construct were observed, the SAR and CFBRs constituted the majority of the negative associations (of course, they also constituted two of the three outcome measures as well). Given that teachers and researchers could have been expecting improved self-regulation for Tools students, the observed negative effects for Tools students on the informant-report measures may be unsurprising.

Nonetheless, it is unlikely that the observed negative effects were purely a product of rater bias for two reasons. First, the abundance of negative effects for the task-based executive function indicators, which do not involve any rater blinding issues, suggests that the negative associations were not purely a product of rater bias. Secondly, the teachers and researchers only knew the children's condition assignment during the pre-kindergarten year, so this potential source of bias would not affect ratings during the kindergarten and first grade assessment follow-ups (i.e., the latent growth rate estimate, many of which were negative for Tools students). Thus, despite the potential for bias among raters, it is unlikely that the lack of blinding fully explains the observed negative associations.

Quality of the Tools activity implementation measures

Although the PRI dataset includes both teacher-reported and researcher-observed activity implementation, both measures exhibit noteworthy shortcomings. Firstly, self-reported measures are inherently biased (Bryman, 2012). In the present study, teachers may have provided inaccurate estimates of their Tools implementation frequency. Some teachers may have inflated their self-reported activity implementation to appease the PRI research team or to avoid reprimand from an administrator. Alternatively, some teachers may have underestimated, overestimated, or simply forgotten their true implementation levels. Any intentional or unintentional distortion by the teacher would also distort the analyses.

As for the researcher-observed activity implementation, the observations occurred during only three days throughout an entire school year (i.e., one day in autumn, one in winter, and one in spring). Thus, the observational data provide a cross-sectional snapshot of implementation; they are not necessarily, or even likely to be, representative of a teachers' typical activity implementation patterns across the school year (Meador, 2015).

Of course, the issue of the observational measure being unrepresentative of teachers' typical implementation patterns is addressed by asking teachers to reflect on their typical implementation patterns through the teacher-reported measure; however, the teacher-reported measure is also biased by the teachers' potentially false memory or rating of his or her own practice. The teacher's bias explains why the researcher-based observational measure is useful to compensate for the relative subjectivity of the teacher-reported implementation measure. Thus, the strengths of one measure were designed to counteract the weaknesses of the other and vice versa.

Although both measures were analyzed in this study so that their strengths would complement one another, the inherent bias in each measure remains a limitation that threatens the study's internal validity (i.e., the measures' bias precludes clear inferences about the observed results). Despite the potential bias in each measure, the results across the two measures were highly consistent, which suggests that the measures' biases did not disproportionately affect the results.

The final limitation with the activity data is that the dataset included implementation data for approximately 45 of the 61 Tools activities. Thus, the results here are only generalizable to the Tools activities that were observed or reported in Tools classrooms as opposed to the full set of Tools activities. Nonetheless, because the teachers were trained in Tools and had access to the Tools manual, it is likely that the activities that were not observed or reported may generally be less central activities in the Tools program. However, without data on the full set of activities, the efficacy of the entire set of Tools practices could not be decisively determined here.

Short-term implementation of the Tools program

In the PRI evaluation study (Farran & Wilson, 2014), students only experienced Tools for one school year. The teachers had also taught Tools the previous year to ensure their familiarity with the program; nonetheless, the outcomes of interest are at the child-level, and the children only had one year of Tools. Given that Tools can be implemented during nursery, preschool, and the kindergarten years, it would have been useful to assess whether sustained Tools implementation over multiple years would be associated with stronger program effects.

Lack of socio-economic status data

As mentioned in section 11.8, because the United States government funded the study (Farran & Wilson, 2014) that gathered these data, United States federal privacy regulations precluded collection of socio-economic data. Once again, all 1140 children in the sample attended public preschool programs aimed at low-income families, which suggests that most of the sample would have low-income backgrounds (Fuhs et al., 2015, p. 210). However, without the relevant data, this assumption cannot be verified.

Consequently, Tools' differential effectiveness across varying socio-economic levels could not be investigated. Several childhood interventions have shown (Blair & Raver, 2014; Tominey & McClelland, 2011; Webster-Stratton et al., 2009; S. J. Wilson & Lipsey, 2007) stronger results in moderation analyses with the low-income segments of the sample, but these sub-group analyses could not be conducted without the requisite data.

13.7.2 Limitations of the analyses

Limited generalizability of the activity analyses

One stated goal of the present study is to identify discrete instructional practices that can be replicated across curricular contexts. That is, even though Tools may not cultivate self-regulation as a wholesale program, Tools may contain individual activities that measurably improve children's self-regulation. Ideally, non-Tools teachers could integrate these specific Tools activities into their practice in hopes of improving student self-regulation.

Despite the intuitive appeal of this logic, the Tools activities that exhibited some positive associations with children's self-regulation (i.e., attention focus, pretend transition, mouse trap, Mr. Wolf, weather graph) may be effective by virtue of their contextualization within a larger curricular framework. That is, although the statistical model isolates the unique effect of the activity on children's self-regulation, teachers still executed that activity within a specific curricular context. Thus, even if this analysis had shown more Tools activities to promote children's self-regulation, the observed effectiveness of those Tools activities could not be assumed to apply across every curricular context.

Equivalence problem in structural equation modeling

The second analytic limitation is that structural equation models, including latent growth models, can prove only that a model fits the data to a certain degree; however, even if the model exhibits perfect fit, this does not imply that the model reflects reality (Blunch, 2008; Maruyama, 1998). In fact, it is possible for other model formulations to fit the data equally well, and that those models may in fact reflect the true relationship among the phenomena (Kline, 2015). Again, only an experimental design whereby children were randomly assigned to various activities could prove true causal relationships among the relevant variables.

In study two, the second research question involved data with both the randomly assigned Tools and comparison groups (total child $n = 1140$). Given the experimental design for these analyses, the structural equation modeling results do illustrate the true causal relationships among the variables (i.e., Tools implementation caused no self-regulation

change at the end of pre-kindergarten and then caused more shallow growth through kindergarten and first grade).

However, for research questions three and four regarding the activity analyses, the structural equation modeling results are still correlational because children were not randomly assigned to different activity types. Given the correlational nature of the results, one could posit alternative formulations of the relationship between Tools activities and self-regulation.

For example, one could speculate that the relationship between make-believe play and self-regulation was reversed in this thesis. Instead, a critic could argue, perhaps Tools teachers implemented more make-believe play among children with lower self-regulation. This would appear as a negative correlation as was observed here (i.e., higher levels of make-believe play associated with lower levels of self-regulation), but the interpretation would be the reverse – teachers would be implementing more make-believe play to support students with lower self-regulation.

In that case, the children's low self-regulation would be causing the teacher to use more make-believe play instead of the other way around. From a Vygotskian lens, this would be an effective pedagogical technique because the teacher would be seeking to improve children's low self-regulation through make-believe play. If that were the case, then the conclusions of this thesis would be backward. That is, it is not the Tools activities that are predicting lower self-regulation but rather lower self-regulation predicting higher Tools implementation by teachers.

There are three key reasons to believe this is not the case. Firstly, although the activity analyses were correlational, the overall curricular evaluation was not. It was a randomized controlled trial whose results indicated lower self-regulation growth for Tools students. Thus, certain elements of the Tools curriculum must be responsible for that relative decline in self-regulation. In the main analysis, the make-believe play, math, and attention-focusing activity blocks all exhibited negative associations with self-regulation, and it is thus likely that those activities partially, or wholly, explain Tools' negative overall effects.

Secondly, the negative associations were observed for both the implementation frequency and fidelity data. Whereas lower self-regulation might compel teachers to implement make-believe play more frequently (i.e., to address children's self-regulatory difficulties), it is less clear how children's low self-regulation would predict higher fidelity from teachers. Moreover, even if children's initially low self-regulation somehow did predict higher fidelity during the pre-kindergarten Tools implementation year, then that higher fidelity should be associated with subsequent self-regulation growth through first grade. This was not the case. In fact, in the analysis where each teacher received a total fidelity score averaged across all activities (see section 12.4.4), higher fidelity did not predict any self-regulation outcomes at the end of pre-kindergarten and instead predicted a significantly lower ($b = -.16, p < .01$) executive function growth rate through kindergarten and first grade.

The hypothesis that higher fidelity predicted lower self-regulation is much more plausible. That is, it is not possible for children's low self-regulation in kindergarten and first grade to cause higher fidelity; this is because the kindergarten self-regulation ratings occurred after the pre-kindergarten fidelity ratings. Causality requires temporal precedence – it is not possible for something in the future to cause something in the past. Thus, the most likely formulation between fidelity and self-regulation would be that high fidelity predicted lower self-regulation, not the reverse.

Thirdly, as a final sensitivity check, I analyzed the inter-correlations among the Tools make-believe play block and children's self-regulation scores at time points one and two. The aim of this robustness check was to investigate whether teachers used more make-believe play among children with higher or lower self-regulation skills. The correlations between Tools play and executive function were significantly positive at time point one ($r = .13, p < .01$) and at time two ($r = .10, p < .05$). This means that Tools teachers reported implementing higher levels of play among children who had higher levels of executive function. The correlations between play and the other two self-regulation variables were non-significant.

Thus, it was not the case that Tools teachers were responding to children's low self-regulation by implementing more make-believe play. Instead, Tools teachers implemented more play among children who already exhibited high self-regulation levels. In the latent growth models with the full vector of covariates where play was included as a predictor of

children's self-regulation development over time, the results showed that higher play implementation and fidelity predicted lower self-regulation growth.

Thus, not only did Tools teachers implement more play among children who already had above average self-regulation, but also higher play implementation subsequently predicted less self-regulation growth for children in those classrooms. Thus, although I cannot decisively prove the causal relationship between Tools activities and self-regulation without an experimental research design, we can be highly confident that the Tools activities predicted lower self-regulation growth and not the reverse.

Potential for Type I error issues

The third strength mentioned in the Strengths section is also a limitation of this study. Specifically, this study is exhaustive because it tested every activity on multiple measures of children's self-regulation. Given the large number of models tested, the Type I error rate also rises; that is, some models may yield illusory statistically significant effects. The model may suggest that an activity significantly affects children's self-regulation, but that finding arises simply by chance.

Although this remains a concern, it is important to note that nearly all the observed significant effects in this study were in the negative direction for Tools students. In fact, in the main analysis with teacher-reported implementation data, only one activity (i.e., weather graphing) exhibited exclusively positive associations with self-regulation outcomes. If Type I error issues were a considerable concern, then we would expect more balance between the significant positive versus negative effects, which was not the case in the present study.

13.8 Directions for future research

The limitations of the present study guide us to valuable directions for future research. Specifically, future research could pursue at least the following three areas to expand upon the work presented here: 1) random assignment to various Tools activity types, 2) a longer implementation period, and 3) more longitudinal follow-ups.

13.8.1 Random assignment to various Tools activity types

Firstly, in the present study, the analyses regarding Tools activities were correlational because students were randomly assigned to either receive the curriculum or not; thus, even if the curriculum had been shown to promote self-regulation, it would not be possible to determine the active ingredients responsible for that growth. Instead, this analysis could only demonstrate whether an activity was significantly *associated* with self-regulation growth rather than whether the activity *caused* that change.

In order to assess a causal link between the activities and self-regulation, it would be necessary to randomly assign students to different activities, or, at least, activity blocks. That way, the effect of activities such as make-believe play could be isolated and measured. In so doing, the Tools developers could determine which parts of the curriculum might be profitably expanded and which might be deemphasized.

Given the large number of Tools activities ($n = 61$), it might be more manageable for future researchers to follow the approach used in this paper to first test activity blocks. For example, instead of testing the four make-believe play activities individually, researchers could test the make-believe play block as a composite whole and quantify its effect on students' self-regulation. That effect could be compared vis-à-vis those students randomly assigned to have Tools without the make-believe play block. By testing the various blocks separately, researchers could identify promising activity blocks (again, through a *causal* framework) in order to inform Tools' future curricular improvement efforts.

13.8.2 Longer implementation period

As described in section 13.7.1, students in the PRI study that generated these data only experienced the Tools curriculum for one year. The Tools curriculum was designed for use in preschool (i.e., three- and four-year-olds), pre-kindergarten (i.e., four- and five-year-olds), and kindergarten (i.e., five- and six-year-olds). As Bronfenbrenner and Morris (2006) argue, “human development takes place through processes of progressively more complex reciprocal interaction between an active, evolving biopsychological human organism and the persons, objects, and symbols in its immediate external environment. To be effective, the interaction must occur on a fairly regular basis over extended periods of time” (p. 797).

Bronfenbrenner's claim that an interaction, or, in this case, an intervention must occur over an extended time period may not be empirically true or feasible given resource constraints, but future research should test this hypothesis in the Tools context. Specifically, a research team could follow a cohort of students through two or three years, as opposed to one year, of Tools implementation to assess whether children's outcomes improve over time.

13.8.3 Longitudinal follow-ups

In addition to extending the implementation period during the early childhood years, a related recommendation for future research is to conduct more longitudinal follow-ups. This is because many early childhood research studies and interventions do not reveal their full effects until years after the study concludes. For example, although Mischel et al.'s (1972) marshmallow study achieved immediate notoriety after publication (Mischel, 2014), the study became increasingly renowned after follow-up studies linked self-regulation to students' SAT scores at age seventeen (Cherniss, 2000) and body mass index ratings at age 30 (Schlam et al., 2013). Similarly, results from the Perry Preschool study (Berrueta-Clement, 1984) have been revisited several decades later to uncover the program's continued cost savings to society (Belfield et al., 2006; Heckman & Masterov, 2007).

Whereas immediate post-test studies can reveal short-term changes in academic achievement and self-regulation, various other life outcomes can only be ascertained years after the study. In the case of this study, Tools produced immediate impacts that neither I nor the curricular developers hypothesized; however, it is also possible that Tools will have long-term effects, either positive or negative, on people's lives that cannot be known at this point in time.

Therefore, it would be useful to follow up a randomly assigned student sample through adolescence and adulthood. While the follow-ups could continue to directly test self-regulation, it might be useful to also test correlates of self-regulation, such as health outcomes, incarceration and drug use rates, marital stability, and other factors that have been associated (Duckworth, 2011; Heckman & Masterov, 2007; Moffitt et al., 2011; Schlam et al., 2013) with self-regulation skills.

13.9 Chapter summary

This chapter described the findings for study two, which investigated Tools' overall effect on self-regulation as well as the associations between Tools activities and self-regulation using a secondary dataset. This chapter also identified study two's strengths and limitations, which then provided guidance toward directions for future research. This dissertation now transitions to the fourteenth and final chapter, which briefly synthesizes the findings from studies one and two and then discusses those findings' implications for policy and practice.

SECTION V: CONCLUSION

CHAPTER 14: Conclusion

This chapter brings the dissertation to a close. In the process, this chapter briefly summarizes the findings from study one and study two. After that, this chapter frames this dissertation's contributions to the existing literature. Finally, this chapter outlines the results' implications for policy and practice.

14.1 Brief summary of study one and two findings

The objective of this dissertation was to evaluate the curricular effects of Tools of the Mind. Given that many researchers (Diamond & Lee, 2011; Hyson et al., 2006) have proclaimed Tools' effectiveness and many schools in the western hemisphere have implemented the program (Bodrova & Leong, 2015a; Turque, 2011), this dissertation aimed to analyze Tools' effectiveness so that educators would have more information regarding its implementation.

To that end, this dissertation contained two studies: First, a research synthesis of Tools, and, second, an analysis of the specific activities that comprise Tools. The results of these two studies are discussed Chapters Ten and Thirteen, but, in summary, students in Tools classrooms did not exhibit the gains hypothesized by the curricular developers (Bodrova & Leong, 2007). Specifically, whereas Tools is hypothesized to significantly improve children's self-regulation, neither the study one nor study two results substantiated this claim.

Rather, the meta-analytic results from study one indicated significant math gains for Tools students versus null effects for self-regulation and literacy. As for the second study, results indicated mostly null and negative associations for Tools activities even after using latent variable models to account for measurement error in the self-regulation construct.

Once again, it is important to contextualize these results in two key ways: Firstly, for the Tools activity analyses, it is important to restate that those data were correlational, which precludes causal claims. That is, we cannot say that Tools activities lowered children's self-regulation scores but rather that higher Tools activity implementation and fidelity were usually associated with lower self-regulation.

The second important factor to contextualize the results is the generally null effects observed across the self-regulation literature. As explained in Chapters Six and Seven, very few, if any, interventions or early childhood curricula can claim to substantially improve children's self-regulation with a variety of samples in a variety of contexts. Although programs such as the Chicago School Readiness Project (Raver et al., 2011) have shown promising preliminary impacts with high-poverty children, no programs have shown consistent self-regulatory gains for mainstream children. Thus, while the Tools' developers claim that Tools significantly improves self-regulation, these results' failure to corroborate that claim is certainly more in alignment with the literature than in contrast to the literature.

14.2 Contributions to the literature base

Although these findings mostly accorded with the existing literature, the present analyses make three novel contributions to the field. First, study one (Section III) includes the first Tools research synthesis study, which incorporated both a systematic review and multilevel meta-analysis. Given the observed variability in existing Tools studies, as well as the gap in results between the published and unpublished Tools literature, the systematic review and meta-analysis provide education policymakers and practitioners with information on the full Tools evidence base. With access to the totality of existing Tools evidence, educators now have more clear guidance regarding whether and how to implement Tools.

The second contribution to the literature is that study two (Section IV) contains the first latent variable structural equation modeling Tools analysis to account for measurement error in the self-regulation construct. That is, whereas previous Tools studies have analyzed raw data from existing self-regulation measures, none of those studies have addressed the pervasive measurement error issue that has been shown (McClelland & Cameron, 2012; Willoughby et al., 2014) to affect much of the self-regulation literature.

This dissertation contains exploratory factor analyses, confirmatory factor analyses, and longitudinal measurement invariance models to confirm the factor structure of the self-regulation constructs over time. By using latent constructs, as opposed to the raw measures, in the analysis, these results provide a less biased view of children's self-regulation skills. It is

hoped that future Tools studies will also employ latent variable modeling to provide a maximally accurate representation of self-regulation change over time.

The third, and final, key contribution to the literature is that study two is the first to examine the associations between specific Tools activities and children's self-regulation. That is, the Tools evaluation studies retrieved for the research synthesis in study one all investigated Tools' effectiveness at the curricular level. None of them analyzed one or more of the specific Tools activities that collectively comprise the curriculum. This dissertation, by contrast, examined all the Tools activities using both teacher-reported and researcher-observed implementation data.

Whereas the dissertation contains several additional findings that represent contributions to the literature (e.g., publication bias in the Tools literature base, various significant associations between a Tools activity and children's self-regulation, etc.), the aforementioned three qualities of this thesis capture three substantial contributions to the literature base. I hope to expand upon these contributions through future research in years to come.

14.3 Implications for policy and practice

In his doctoral thesis, Bronfenbrenner (1979) described the connection between academic research and public policy, arguing that we need "not merely a complementary relation between these two domains but rather their functional integration" (p. 8). In keeping with Bronfenbrenner's logic, this dissertation aimed to address a question that would be directly relevant for both research and policy in hopes of more effectively connecting these domains.

Specifically, this dissertation evaluated Tools' curricular effects in order to inform policymakers' and practitioners' implementation of the program. Based on the findings presented herein, it seems unwise to continue rapid expansion of Tools implementation. The meta-analysis in study one belied the claim that Tools improves children's self-regulation relative to other early childhood curricula. Although Tools exhibited significant effects for children's math skills, the Tools developers themselves identified math as an area of weakness for Tools (Mackay, 2013). Thus, for those seeking to promote children's self-regulation skills, the meta-analytic results do not support Tools implementation.

Moreover, the study two analyses revealed that the curriculum as a whole, as well as its constituent activities, mostly exhibited either null or negative associations with children's self-regulation. The effect was consistent when investigating both teacher-reported implementation frequency as well as researcher-observed implementation fidelity.

Of course, just because Tools did not consistently outperform the comparator curricula does not necessarily signify Tools' inferiority; Tools may simply be no better than, or equally as good as, many existing curricula. However, given Tools' implementation costs of \$3000 per classroom per year (United States Department of Education, 2008b), which in Washington DC required a \$1.5 million increase in the city's curriculum budget (Turque, 2011), resource-poor schools may wish to carefully consider the allocation of their curricular finances.

In addition to investigating Tools' effectiveness at the curricular level, study two of this dissertation also aimed to identify discrete instructional practices that could be replicated within and beyond the Tools context. However, with the exception of the weather graphing activity (i.e., where teachers and children chart the daily weather patterns on a class graph), none of the other tested activities consistently predicted improved self-regulation. Thus, it would be irresponsible to direct teachers to incorporate specific Tools activities into their instructional practice without more evidence of these activities' effectiveness.

Despite these largely null and negative findings in studies one and two, it would also be premature to abandon the Tools approach completely. Tools has shown promising results in multiple studies (Barnett et al., 2008; Blair & Raver, 2014; Diamond et al., 2007), and schools across North America and now South America have adopted the program (Behrman, 2016; Turque, 2011). Thus, it is possible that Tools can work in certain situations for certain teachers and students.

One aim of this dissertation had been to identify the specific Tools activities that might drive the curriculum forward, as well as identify specific student subgroups and contexts for which Tools is especially effective. Although this dissertation yielded mostly null and negative results for Tools activities, it remains possible that Tools could be modified in a small-scale setting to benefit those who implement it.

In a similar vein, two of the most well-known and most successful preschool studies, the Perry Preschool project (Berrueta-Clement, 1984) and the Abdecarian project (Ramey & Campbell, 1984), were implemented with a small number of students and highly trained teachers (Hyson et al., 2006). Those programs lack evidence of effective scaling among large numbers of students or proliferation across multiple geographic settings, but their effectiveness on a small scale remains compelling (Heckman et al., 2010).

Thus, perhaps Tools should wait to expand until it has refined the fundamental elements of its curriculum and better determined how to train participating teachers in the method. This is because Tools “requires a fairly high level of theoretical and pedagogical insight” and “assumes a high degree of buy-in or commitment on the part of teachers” (Ryan, 2004, p. 671), which complicates the prospects for effective scaling.

Nonetheless, given the promising preliminary results from various Tools RCT studies (Barnett et al., 2008; Blair & Raver, 2014; Diamond et al., 2007), it seems possible that, as with the Perry preschool and Abdecarian projects, the program can work well when implemented and tested on a smaller scale. Thus, looking forward, perhaps the Tools developers should lead a small-scale Tools implementation study that identifies the core components of the program as well as the optimal training regimen for teachers.

In sum, this dissertation illustrated both the potential and the observed shortcomings of the Tools program. If the developers can implement Tools so that the program consistently delivers on its stated outcomes, then Tools will be among the first early childhood curricula to improve children’s self-regulation, and, thus, improve children’s lives.

References

- Ainsworth, M. S. (1979). Infant-mother attachment. *American Psychologist*, 34(10), 932.
- Alexander, K., Entwisle, D., & Kabbani, N. (2001). The dropout process in life course perspective: Early risk factors at home and school. *The Teachers College Record*, 103(5), 760–822.
- Alloway, T., Gathercole, S., & Pickering, S. (2006). Verbal and visuospatial short-term and working memory in children: Are they separable? *Child Development*, 77(6), 1698–1716.
- Alloway, T., Lawrence, A., & Rodger, S. (2013). Antisocial behavior: Exploring behavioral, cognitive, and environmental influences on expulsion. *Applied Cognitive Psychology*, 27(4), 520–526.
- Anokhin, A. P., Heath, A. C., & Myers, E. (2004). Genetics, prefrontal cortex, and cognitive control: A twin study of event-related brain potentials in a response inhibition task. *Neuroscience Letters*, 368(4), 314–318.
- Arda, T. B. (2012). Social competence and promoting alternative thinking strategies: PATHS preschool curriculum. *Theory Into Practice*, 12(4), 2691–2698.
- Baker-Henningham, H., Walker, S., Powell, C., & Gardner, J. M. (2009). A pilot study of the Incredible Years Teacher Training programme and a curriculum unit on social and emotional skills in community pre-schools in Jamaica. *Childcare, Health and Development*, 35, 624–631.
- Balshem, H., Helfand, M., Schünemann, H. J., Oxman, A. D., Kunz, R., Brozek, J., & Guyatt, G. H. (2011). GRADE guidelines: Rating the quality of evidence. *Journal of Clinical Epidemiology*, 64(4), 401–406.
- Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 84(2), 191–215.

- Bandura, A. (1991). Social cognitive theory of self-regulation. *Organizational Behavior and Human Decision Processes*, 50(3), 248–287.
- Barker, J. E., Semenov, A. D., Michaelson, L., Provan, L. S., Snyder, H. R., & Munakata, Y. (2014). Less-structured time in children’s daily lives predicts self-directed executive functioning. *Frontiers in Psychology*, 5, 1–16.
- Barkley, R. A., Shelton, T. L., Crosswait, C., Moorehouse, M., Fletcher, K., Barrett, S., & Metevia, L. (2000). Multi-method psycho-educational intervention for preschool children with disruptive behavior: Preliminary results at post-treatment. *Journal of Child Psychology and Psychiatry*, 41(3), 319–332.
- Barnett, W. S., Jung, K., Yarosz, D. J., Thomas, J., Hornbeck, A., Stechuk, R., & Burns, S. (2008). Educational effects of the Tools of the Mind curriculum: A randomized trial. *Early Childhood Research Quarterly*, 23(3), 299–313.
- Baron, A. (2014). *Parenting practices predictive of self-regulation development in young children: A secondary data analysis of the Millennium Cohort Study (Unpublished master’s thesis)*. University of Oxford.
- Baron, A., & Chen, H. L. S. (2012). Looking in a science classroom: Exploring possibilities of creative cultural divergence in science teaching and learning. *Culture Studies of Science Education*, 7(1), 93–101.
- Baron, A., Evangelou, M., Malmberg, L. E., & Melendez-Torres, G. J. (2016). Protocol for a systematic review: The Tools of the Mind curriculum for improving self-regulation in early childhood: A systematic review. Philadelphia, Pennsylvania: Campbell Collaboration.
- Baumrind, D., Larzelere, R. E., & Owens, E. B. (2010). Effects of preschool parents’ power assertive patterns and practices on adolescent development. *Parenting: Science and Practice*, 10(3), 157–201.
- Behrman, J. (2016). Data request for Tools Of The Mind review (email). University of

Pennsylvania.

- Belfield, C. R., Nores, M., Barnett, S., & Schweinhart, L. (2006). The High/Scope Perry preschool program: Cost-benefit analysis using data from the age-40 follow-up. *Journal of Human Resources*, 41(1), 162–190. <http://doi.org/10.3368/jhr.XLI.1.162>
- Bell, M. A., & Deater-Deckard, K. (2007). Biological systems and the development of self-regulation: Integrating behavior, genetics, and psychophysiology. *Journal of Developmental & Behavioral Pediatrics*, 28(5), 409–420. <http://doi.org/10.1097/DBP.0b013e3181131fc7>
- Ben-Ari, E. (1995). Forms of quality? Documentation and discipline in a Japanese preschool. *Excellence and Quality in Education*, 32, 61–81.
- Berk, L. E., & Meyers, A. B. (2013). The role of make-believe play in the development of executive function: Status of research and future directions. *American Journal of Play*, 6(1), 98–110.
- Bernier, A., Carlson, S. M., & Whipple, N. (2010). From external regulation to self-regulation: Early parenting precursors of young children's executive functioning. *Child Development*, 81(1), 326–339. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/20331670>
- Berrueta-Clement, J. (1984). Changed lives: The effects of the Perry preschool program on youths through age 19. *Monographs of the High/Scope Educational Research Foundation*, 393(6), 1–193. Retrieved from <http://files.eric.ed.gov/fulltext/ED313128.pdf>
- Berthoud, R., & Gershuny, J. (Eds.). (2000). *Seven years in the lives of British families: Evidence of the dynamics of social change from the British Panel Survey*. London: Policy Press.
- Biederman, J., Hirshfeld-Becker, D., & Rosenbaum, J. (2001). Further evidence of association between behavioral inhibition and social anxiety in children. *American Journal of Psychiatry*, 158, 1673–1679.
- Bierman, K. L., Domitrovich, C., Blair, C., Nelson, K., & Gill, S. (2008). Promoting academic and social-emotional school readiness: The Head Start REDI program. *Child*

Development, 79(6), 1802–1817. Retrieved from
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3549580&tool=pmcentrez&rendertype=abstract>

Bierman, K. L., Nix, R. L., & Greenberg, M. T. (2008). Executive functions and school readiness intervention: Impact, moderation, and mediation in the Head Start REDI program. *Development and Psychopathology*, 20(3), 821–843.

Bierman, K., & Torres, M. (2016). Promoting the development of executive functions through early education and prevention programs. In J. Griffin, L. Freund, & P. McCardle (Eds.), *Executive function in preschool age children: Integrating measurement, neurodevelopmental and translational research* (Vol. 1, pp. 35–50). Washington DC: American Psychological Association. <http://doi.org/10.1002/cd.234>

Blackwell, K., Cepeda, N., & Munakata, Y. (2009). When simple things are meaningful: Working memory strength predicts children's cognitive flexibility. *Journal of Experimental Child Psychology*, 103(2), 241–249. Retrieved from
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2737814&tool=pmcentrez&rendertype=abstract>

Blair, C. (2003). Behavioral inhibition and behavioral activation in young children: Relations with self-regulation and adaptation to preschool in children attending Head Start. *Developmental Psychobiology*, 42, 301–311.

Blair, C. (2006). How similar are fluid cognition and general intelligence? A developmental neuroscience perspective on fluid cognition as an aspect of human cognitive ability. *Behavioral and Brain Sciences*, 29(2), 109–125. Retrieved from
<http://www.ncbi.nlm.nih.gov/pubmed/16606477>

Blair, C., & Diamond, A. (2008). Biological processes in prevention and intervention: The promotion of self-regulation as a means of preventing school failure. *Development and Psychopathology*, 20(3), 899–911. Retrieved from
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2593474&tool=pmcentrez&rendertype=abstract>

- Blair, C., & Raver, C. (2014). Closing the achievement gap through modification of neurocognitive and neuroendocrine function: Results from a cluster randomized controlled trial of an innovative approach to the education of children in kindergarten. *PloS One*, 9(11), 1–13.
- Blair, C., & Raver, C. (2015). School readiness and self-regulation: A developmental psychobiological approach. *Annual Review of Psychology*, 66(1), 711–731. Retrieved from <http://www.annualreviews.org/doi/abs/10.1146/annurev-psych-010814-015221>
- Blair, C., & Razza, R. (2007). Relating effortful control, executive function, and false belief understanding to emerging math and literacy ability in kindergarten. *Child Development*, 78(2), 647–663. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/17381795>
- Blair, C., & Ursache, A. (2011). A bidirectional model of executive functions and self-regulation. In K. Vohs & R. Baumeister (Eds.), *Handbook of self-regulation: Research, theory, and applications* (2nd ed., pp. 300–320). New York: Guilford Press.
- Blunch, N. (2008). *Introduction to structural equation modeling using SPSS and AMOS*. London: Sage.
- Bodrova, E., & Leong, D. (2001). Tools of the Mind: A case study of implementing the Vygotskian approach in American early childhood and primary classrooms. *Innodata Monographs*, 7, 1–44. Retrieved from <http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=ED455014&site=ehost-live>
- Bodrova, E., & Leong, D. (2003). Chopsticks and counting chips: Do play and foundational skills need to compete for the teacher's attention in an early childhood classroom? *Young Children*, 5, 1–8.
- Bodrova, E., & Leong, D. (2005). Self-regulation: A foundation for early learning. *Principal*, 85(1), 30–35.
- Bodrova, E., & Leong, D. (2007). *Tools of the Mind: The Vygotskian approach to early childhood*

education (2nd ed.). Columbus: Pearson.

Bodrova, E., & Leong, D. (2008). Developing self-regulation in young children: Can we keep all the crickets in the basket? *Young Children*, 63(2), 56–58.

Bodrova, E., & Leong, D. (2013). Play and self-regulation: Lessons from Vygotsky. *American Journal of Play*, 6(1), 111–123.

Bodrova, E., & Leong, D. (2015a). Locations and alignments. Retrieved January 7, 2014, from <http://toolsofthemind.org/learn/locations-alignments/>

Bodrova, E., & Leong, D. (2015b). Tools of the Mind. Retrieved January 14, 2015, from <http://toolsofthemind.org>

Bodrova, E., Leong, D., & Akhutina, T. (2011). When everything new is well-forgotten old: Vygotsky/Luria insights in the development of executive functions. *New Directions for Child and Adolescent Development*, 133, 11–28.

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, 120(3), 322–330. Retrieved from <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3369499&tool=pmcentrez&rendertype=abstract>

Bor, W., Sanders, M. R., & Markie-Dadds, C. (2002). The effects of the Triple P: Positive Parenting Program on preschool children with co-occurring disruptive behavior and attentional/hyperactive difficulties. *Journal of Abnormal Child Psychology*, 30(6), 571–587.

Boyer, W. (2012). Getting back to the woods: Familial perspectives on culture and preschoolers' acquisition of self-regulation and emotion regulation. *Early Childhood Education Journal*, 41(2), 153–159. Retrieved from <http://link.springer.com/10.1007/s10643-012-0536-7>

Bronfenbrenner, U. (1974). Developmental research, public policy, and the ecology of childhood. *Child Development*, 45, 1–5.

- Bronfenbrenner, U. (1979). *The ecology of human development: Experiments by nature and design*. Cambridge, MA: Harvard University Press.
- Bronfenbrenner, U. (1986). Recent advances in research on the ecology of human development. In R. K. Silbereisen, K. Eyferth, & G. Rudinger (Eds.), *Development as action in context: Problem behavior and normal youth development* (pp. 286–309). New York: Springer-Verlag.
- Bronfenbrenner, U. (1989). The ecology of the family as a context for human development: Research perspectives. *Developmental Psychology*, 22(1), 723–742.
- Bronfenbrenner, U., & Ceci, S. (1994). Nature-nurture reconceptualized in developmental perspective: A bioecological model. *Psychological Review*, 101(4), 568–586.
<http://doi.org/doi:10.1037/0033-295X.101.4.568>
- Bronfenbrenner, U., & Morris, P. (2006). The bioecological model of human development. In *Handbook of Child Psychology* (pp. 793–828). Hoboken: Wiley.
- Bronson, M. (2000). *Self-regulation in early childhood: Nature and nurture*. New York: Guilford Press.
- Bryman, A. (2012). *Social research methods*. New York: Oxford University Press.
- Bryne, B. (2012). *Structural equation modeling with Mplus*. London: Routledge.
- Bull, R., & Scerif, G. (2010). Executive functioning as a predictor of children's mathematics ability: Inhibition, switching, and working memory. *Developmental Neuropsychology*, 19(3), 273–293.
- Bushman, B. J., & Wells, G. L. (2001). Narrative impressions of literature: The availability bias and the corrective properties of meta-analytic approaches. *Personality and Social Psychology Bulletin*, 27(9), 1123–1130.
- Calkins, S. D., Howse, R. B., & Philippot, P. (2004). Individual differences in self-regulation: Implications for childhood adjustment. In *The regulation of emotion* (pp. 307–332).

- Calvin, C. M., Deary, I. J., Webbink, D., Smith, P., & Visscher, P. M. (2012). Multivariate genetic analyses of cognition and academic achievement from two population samples of 174,000 and 166,000 school children. *Behavior Genetics*, 42(5), 699–710.
- Cameron, C. E., Brock, L. L., Murrah, W. M., Bell, L. H., Worzalla, S. L., Grissmer, D., & Morrison, F. J. (2012). Fine motor skills and executive function both contribute to kindergarten achievement. *Child Development*, 83(4), 1229–1244.
- Camp, C. J., Judge, K. S., Bye, C. A., Fox, K. M., & Mattern, J. M. (1997). An intergenerational program for persons with dementia using Montessori methods. *The Gerontologist*, 37(5), 688–692.
- Carlson, S. (2003). Executive function in context: Development, measurement, theory, and experience. *Monographs of the Society for Research in Child Development*, 68, 138–151.
- Carlson, S. M., White, R. E., & Davis-Unger, A. (2014). Evidence for a relation between executive function and pretense representation in preschool children. *Cognitive Development*, 29, 1–16. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/24357896>
- Carlton, P. L., & Strawderman, W. E. (1996). Evaluating cumulated research: The inadequacy of traditional methods. *Biological Psychiatry*, 39(1), 65–72.
- Center for Disease Control. (2013). *Trends in the parent-report of health care provider-diagnosis and medication treatment for ADHD: United States, 2003-2011*. Atlanta, Georgia.
- Chan, A. W., Hróbjartsson, A., Haahr, M. T., Gøtzsche, P. C., & Altman, D. G. (2004). Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. *Journal of American Medical Association*, 291(20), 2457–2465.
- Charlesworth, R., Hart, C. H., Burts, D. C., Thomasson, R. H., Mosley, J., & Fleege, P. O. (1993). Measuring the developmental appropriateness of kindergarten teachers' beliefs and practices. *Early Childhood Research Quarterly*, 8(3), 255–276.

- Cherniss, C. (2000). Emotional intelligence: What it is and why it matters. In *Society for Industrial and Organizational Psychology*. New Orleans, LA.
- Chularut, P., & DeBacker, T. K. (2004). The influence of concept mapping on achievement, self-regulation, and self-efficacy in students of English as a second language. *Contemporary Educational Psychology*, 29(3), 248–263.
- Clements, D., & Sarama, J. (2012). The efficacy of an intervention synthesizing scaffolding designed to promote self-regulation with an early mathematics curriculum: Effects on executive function. In *Society for research on educational effectiveness* (pp. 1–7).
- Clements, D., Sarama, J., & Germeroth, C. (2016). Learning executive function and early mathematics: Directions of causal relations. *Early Childhood Research Quarterly*, 36, 79–90. <http://doi.org/10.1016/j.ecresq.2015.12.009>
- Clements, D., Sarama, J., Germeroth, C., & Fesler, L. (2014). Effects on executive function and mathematics learning of an early mathematics curriculum synthesized with scaffolded play designed to promote self-regulation versus the mathematics curriculum alone (in press). *Cognition and Instruction*, 1–67.
- Cooper-Kahn, J., & Foster, M. (2013). *Boosting executive skills in the classroom: A practical guide for educators*. San Francisco: Jossey-Base.
- Cooper, D., & Farran, D. (1988). Behavioral risk in kindergarten. *Early Childhood Research Quarterly*, 3, 1–20.
- Corsi, P. (1972). Human memory and the medial temporal regions of the brain. *Dissertation Abstract International*, 34(2), 891.
- Creative curriculum solutions website. (2015). Retrieved January 26, 2015, from <http://teachingstrategies.com/curriculum/>
- Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Science*, 13(4), 148–153.
- Daly, M., Delaney, L., & Baumeister, R. F. (2015). Childhood self-control and

unemployment throughout the life course: Evidence from two British cohort studies. *Psychological Science*, 26(6), 709–723. <http://doi.org/0956797615569001>

Daunic, A. P., Smith, S. W., Garvan, C. W., Barber, B. R., & Becker, M. K. (2012). Reducing developmental risk for emotional/behavioral problems: A randomized controlled trial examining the Tools for Getting Along curriculum. *Journal of School Psychology*, 50(2), 149–166.

David, P., Müller, U., Frye, D., Marcovitch, S., Carlson, S. M., & Zelazo, P. D. (2003). The development of executive function in early childhood. *Monographs of the Society for Research in Child Development*, 68(3), 1–158.

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, 125(6), 627–668. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/10589297>

Deeks, J., & Higgins, J. (2010). *Statistical algorithms in Review Manager 5 on behalf of the Statistical Methods Group of The Cochrane Collaboration. Statistical Methods Group.*

Denham, S., Warren-Khot, H., Bassett, H. H., Wyatt, T., & Perna, A. (2012). Factor structure of self-regulation in preschoolers: Testing models of a field-based assessment for predicting early school readiness. *Journal of Experimental Child Psychology*, 111(3), 386–404.

Dennis, T. (2006). Emotional self-regulation in preschoolers: The interplay of child approach reactivity, parenting, and control capacities. *Developmental Psychology*, 42(1), 84–97. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/16420120>

Dewey, J. (1915). *School and society*. Chicago: University of Chicago Press.

Diamond, A. (2006). The early development of executive functions. In E. Bialystok & F. Craik (Eds.), *Lifespan cognition: Mechanisms of change* (pp. 70–95). London: Oxford University Press.

Diamond, A. (2012). Activities and programs that improve children's executive functions.

Current Directions in Psychological Science, 21(5), 335–341. Retrieved from
<http://cdp.sagepub.com/lookup/doi/10.1177/0963721412453722>

- Diamond, A., Barnett, S., Thomas, J., & Munro, S. (2007). Preschool program improves cognitive control. *Science*, 318, 1387–1388.
- Diamond, A., & Lee, K. (2011). Interventions shown to aid executive function development in children 4-12 years old. *Science*, 333, 959–964.
- Diaz, R. M., & Berk, L. E. (Eds.). (2014). *Private speech: From social interaction to self-regulation*. Oxford: Psychology Press.
- District of Columbia Public Schools. (2016). Early childhood academic programs. Retrieved January 11, 2016, from <http://dcps.dc.gov/page/early-childhood-academic-programs>
- Dix, T. (1991). The affective organization of parenting: Adaptive and maladaptive processes. *Psychological Bulletin*, 110(1), 3–25.
- Dodge, D. T., Colker, L. J., & Heroman, C. (2001). *A teacher's guide to using the Creative Curriculum developmental continuum assessment system*. Bethesda, Maryland.
- Domitrovich, C. E., Cortes, R. C., & Greenberg, M. T. (2007). Improving young children's social and emotional competence: A randomized trial of the preschool "PATHS" curriculum. *Journal of Primary Prevention*, 28(2), 67–91.
- Doyle, J., Waters, E., Noonan, E., Royle, N., Kowalski, K., Gill, C., & Wilson, D. B. (2010). Working with The Campbell Collaboration to produce reviews of relevance to public health across the education, justice and social welfare sectors. *Journal of Public Health*, 1–3.
- Duckworth, A. (2011). The significance of self-control. *Proceedings of the National Academy of Sciences of the United States of America*, 108(7), 2639–2640.
<http://doi.org/10.1073/pnas.1019725108>
- Duckworth, A., & Carlson, S. (2013). Self-regulation and school success. In B. Sokol (Ed.),

Self-regulation and autonomy (pp. 208–230). Cambridge: Cambridge University Press.

Duckworth, A., & Kern, M. (2012). A meta-analysis of the convergent validity of self-control measures. *Journal of Personality and Social Psychology*, 45(3), 259–268.

Duckworth, A., Kim, B., & Tsukayama, E. (2013). Life stress impairs self-control in early adolescence. *Frontiers in Psychology*, 3, 1–12. <http://doi.org/10.3389/fpsyg.2012.00608>

Duckworth, A., & Seligman, M. (2005). Self-discipline outdoes IQ in predicting academic performance of adolescents. *Psychological Science*, 16(12), 939–944.

Duckworth, A., & Steinberg, L. (2015). Unpacking self-control. *Child Development Perspectives*, 9(1), 32–37. <http://doi.org/10.1111/cdep.12107>

Duncan, G., Dowsett, C., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., & Japel, C. (2007). School readiness and later achievement. *Developmental Psychology*, 43(6), 1–36.

Duncan, T., & Duncan, S. (2009). The ABC's of LGM: An introductory guide to latent variable growth curve modeling. *Social and Personality Psychology Compass*, 3(6), 979–991. <http://doi.org/10.1111/j.1751-9004.2009.00224.x>.The

Duncan, T., Duncan, S., & Strycker, L. (2006). *An introduction to latent variable growth curve modeling* (2nd ed.). Mahwah: Lawrence Erlbaum Associates.

Eisenberg, N., Valiente, C., Morris, A., Gershoff, E., & Shepard, S. (2003). Longitudinal relations among parental emotional expressivity, children's regulation, and quality of socioemotional functioning. *Developmental Psychology*, 39(1), 3–19. Retrieved from <http://doi.apa.org/getdoi.cfm?doi=10.1037/0012-1649.39.1.3>

Eisenberg, N., Zhou, Q., Spinrad, T. L., Valiente, C., Fabes, R. A., & Liew, J. (2005). Relations among positive parenting, children's effortful control, and externalizing problems: A three-wave longitudinal study. *Child Development*, 76(5), 1055–1071.

Elias, C. L., & Berk, L. E. (2002). Self-regulation in young children: Is there a role for

sociodramatic play? *Early Childhood Research Quarterly*, 17(2), 216–238.

Elkonin, D. (1975). Primary schoolchildren's intellectual capabilities and the content of instruction. *Soviet Studies in the Psychology of Learning and Teaching Mathematics*, 7, 13–54.

Elliott, D. S., & Mihalic, S. (2004). Issues in disseminating and replicating effective prevention programs. *Prevention Science*, 5(1), 47–53.

Ellis, P. (2010). *The essential guide to effect sizes: An introduction to statistical power, meta-analysis and the interpretation of research results*. Cambridge: Cambridge University Press.

Enders, C. (2010). *Applied missing data analysis*. New York: Guilford Press.

Evans, G. W. (2003). A multimethodological analysis of cumulative risk and allostatic load among rural children. *Developmental Psychology*, 39(5), 924–933. Retrieved from <http://doi.apa.org/getdoi.cfm?doi=10.1037/0012-1649.39.5.924>

Evans, G. W., Kim, P., Ting, A. H., Tesher, H. B., & Shannis, D. (2007). Cumulative risk, maternal responsiveness, and allostatic load among young adolescents. *Developmental Psychology*, 43(2), 341.

Fantuzzo, J., Manz, P., & McDermott, P. (1998). Preschool version of the social skills rating system: An empirical analysis of its use with low-income children. *Journal of School Psychology*, 36(2), 199–214.

Farran, D., & Wilson, S. (2014). Achievement and self-regulation in pre-kindergarten classrooms: Effects of the Tools of the Mind curriculum (in press). *Child Development*, 1–66.

Farran, D., Wilson, S., Meador, D., Norvell, J., & Nesbitt, K. (2015). Experimental evaluation of the Tools of the Mind pre-K curriculum technical report. Nashville, TN: Peabody Research Institute. Retrieved from <https://my.vanderbilt.edu/toolsofthemindevaluation/>

Feldman, R., Eidelman, A. I., & Rotenberg, N. (2014). Parenting stress, infant emotion

regulation, maternal sensitivity, and the cognitive development of triplets: A model for parent and child influences in a unique ecology. *Child Development*, 75(6), 1774–1791.

Fernyhough, C. (2009). Vygotsky, Luria, and the Social Brain. In B. Sokol, U. Müller, J. Carpendale, A. Young, & G. Iarocci (Eds.), *Self- and social-regulation: Exploring the relations between social interaction, social cognition, and the development of executive functions* (pp. 1–38). Oxford: Oxford University Press.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics: And sex and drugs and rock “n” roll* (4th ed.). London: Sage.

Fitzpatrick, C., McKinnon, R., Blair, C., & Willoughby, M. (2014). Do preschool executive function skills explain the school readiness gap between advantaged and disadvantaged children? *Learning and Instruction*, 30, 25–31. Retrieved from <http://linkinghub.elsevier.com/retrieve/pii/S0959475213000819>

Flouri, E., Midouhas, E., & Joshi, H. (2014). Family poverty and trajectories of children’s emotional and behavioural problems: The moderating roles of self-regulation and verbal cognitive ability. *Journal of Abnormal Child Psychology*, 42(1), 1–14.

Flouri, E., Tzavidis, N., & Kallis, C. (2010). Area and family effects on the psychopathology of the Millennium Cohort Study children and their older siblings. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 51(2), 152–61. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/19804382>

Ford, T., Edwards, V., Sharkey, S., Ukoumunne, O., Byford, S., Norich, B., & Logan, S. (2012). Supporting teachers and children in schools: The effectiveness and cost-effectiveness of the Incredible Years teacher classroom management programme in primary school children. *BMC Public Health*, 12(1), 719.

Fox, E., & Riconscente, M. (2008). Metacognition and self-regulation in James, Piaget, and Vygotsky. *Educational Psychology Review*, 20, 373–389.

Freud, S. (1899). *The interpretation of dreams*. New York: Oxford University Press.

- Friedman, N., Miyake, A., Young, S. E., Defries, J. C., Corley, R. P., & Hewitt, J. K. (2008). Individual differences in executive functions are almost entirely genetic in origin. *Journal of Experimental Psychology. General*, 137(2), 201–225. <http://doi.org/10.1037/0096-3445.137.2.201>
- Fuhs, M., Farran, D., & Nesbitt, K. (2013). Preschool classroom processes as predictors of children's cognitive self-regulation skills development. *School Psychology Quarterly*, 28(4), 347–359. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/23895316>
- Fuhs, M., Farran, D., & Nesbitt, K. (2015). Prekindergarten children's executive functioning skills and achievement gains: The utility of direct assessments and teacher ratings. *Journal of Educational Psychology*, 107(1), 207–221.
- Fuhs, M., Nesbitt, K., Farran, D., & Dong, N. (2014). Longitudinal associations between executive functioning and academic skills across content areas. *Developmental Psychology*, 50(6), 1698–1709. <http://doi.org/10.1037/a0036633>
- Gathercole, S. E., Pickering, S. J., Knight, C., & Stegmann, Z. (2004). Working memory skills and educational attainment: Evidence from national curriculum assessments at 7 and 14 years of age. *Applied Cognitive Psychology*, 18(1), 1–16. Retrieved from <http://doi.wiley.com/10.1002/acp.934>
- Gellis, Z., & Reid, W. J. (2004). Strengthening evidence-based practice. *Brief Treatment and Crisis Intervention*, 4(2), 155.
- Gershoff, E. T. (2002). Corporal punishment by parents and associated child behaviors and experiences: A meta-analytic and theoretical review. *Psychological Bulletin*, 128(4), 539–579. Retrieved from <http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-2909.128.4.539>
- Gilliam, W. S., & Zigler, E. F. (2004). A critical meta-analysis of all evaluations of state-funded preschool from 1977 to 1998: Implications for policy, service delivery and program evaluation. *Early Childhood Research Quarterly*, 15(4), 441–473.
- Gilliam, W. S. (2005). *Prekindergarteners left behind: Expulsion rates in state prekindergarten systems*.

New York: Foundation for Child Development.

- Gilmore, C., Attridge, N., Simms, V., & Inglis, M. (2013). Individual differences in inhibitory control, not non-verbal number acuity, correlate with mathematics achievement. *Plos*, 8(6), 1–9. Retrieved from <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3681957&tool=pmcentrez&rendertype=abstract>
- Gioia, G., Espy, K. A., & Isquith, P. K. (2005). Behavior Rating Inventory of Executive Function—Preschool Version (BRIEF-P). Odessa, Florida: Psychological Assessment Resources.
- Glaser, B. G. (1963). Retreading research materials: The use of secondary data analysis by the independent researcher. *American Behavioral Scientist*, 6(11), 11–14.
- Glaser, D. (2000). Child abuse and neglect and the brain: A review. *Journal of Child Psychology and Psychiatry*, 41(1), 97–116. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/10763678>
- Goldstein, H. (2011). *Multilevel statistical models*. Oxford: Wiley and Sons.
- Goldstein, H., Yang, M., Omar, R., Turner, R., & Thompson, S. (2000). Meta-analysis using multilevel models with an application to the study of class size effects. *Applied Statistics*, 49(3), 399–412. <http://doi.org/10.1111/1467-9876.00200>
- Goldstein, S., & Brooks, R. (2007). Problems of self-discipline. In *Understanding and managing children's classroom behavior: Creating sustainable, resilient classrooms* (2nd ed., pp. 113–127). Hoboken: Wiley and Sons.
- Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, 60, 549–576.
- Gredler, M. E. (2009). Hiding in plain sight: The stages of mastery/self-regulation in Vygotsky's cultural-historical theory. *Educational Psychologist*, 44, 1–19.

- Greenberg, M., Kusche, C., & Cook, E. (1995). Promoting emotional competence in school-aged children: The effects of the PATHS curriculum. *Development and Psychopathology*, 7(1), 117–136.
- Grolnick, W. S., & Ryan, R. M. (1989). Parent styles associated with children's self-regulation and competence in school. *Journal of Educational Psychology*, 81(2), 143–154. Retrieved from <http://doi.apa.org/getdoi.cfm?doi=10.1037/0022-0663.81.2.143>
- Gulchak, D. J. (2008). Using a mobile handheld computer to teach a student with an emotional and behavioral disorder to self-monitor attention. *Education and Treatment of Children*, 31(4), 567–581.
- Guyatt, G. H., Oxman, A. D., Schünemann, H. J., Tugwell, P., & Knottnerus, A. (2011). GRADE guidelines: A new series of articles in the Journal of Clinical Epidemiology. *Journal of Clinical Epidemiology*, 64(4), 380–382.
- Hammer, C., Blair, C., Lopez, L., Leong, D., & Bodrova, E. (2012). Tools of the Mind: Promoting the school readiness of ELLs. In *Society for Research on Educational Effectiveness* (pp. 1–6). Washington DC.
- Hannafin, R. D. (2004). Achievement differences in structured versus unstructured instructional geometry programs. *Educational Technology Research and Development*, 52(1), 19–32. Retrieved from <http://link.springer.com/10.1007/BF02504770>
- Harris, K. R., Friedlander, B. D., & Graham, S. (2005). Self-monitoring of attention versus self-monitoring of academic performance: Effects among students with ADHD in the general education classroom. *The Journal of Special Education*, 39(3), 145–156.
- Heatherton, T. F., & Wagner, D. D. (2012). Cognitive neuroscience of self-regulation failure. *Trends in Cognitive Science*, 15(3), 132–139.
- Heckman, J., & Masterov, D. (2007). The productivity argument for investing in young children. *Review of Agricultural Economics*, 29(3), 446–493. Retrieved from <http://www.nber.org/papers/w13016.pdf>

- Heckman, J., Moon, S. H., Pinto, R., Savelyev, P. A., & Yavitz, A. (2010). The rate of return to the HighScope Perry Preschool Program. *Journal of Public Economics*, 94(1), 114–128.
- Hickey, G., McGilloway, S., Donnelly, M., & O'Neill, D. (2015). Exploring the effects of a universal classroom management training programme on teacher and child behaviour: A group randomised controlled trial and cost analysis. *Journal of Early Childhood Research*, 1–21. <http://doi.org/10.1177/1476718X15579747>
- Higgins, J., & Green, S. (Eds.). (2011). *Cochrane handbook for systematic reviews of interventions*. Cochrane Collaboration. Retrieved from <http://handbook.cochrane.org/>
- Highfield, R. (2007, November 29). “Formal play” better prepares children for school. *The Telegraph*, pp. 1–2. London. Retrieved from <http://www.telegraph.co.uk/news/science/science-news/3316565/Formal-play-better-prepares-children-for-school.html>
- HighScope: Inspiring educators to inspire children. (2015). Retrieved October 17, 2015, from <http://www.highscope.org/Content.asp?ContentId=761>
- Hofferth, S., & Sandberg, J. (2001). How American children spend their time. *Journal of Marriage and Family*, 63(5), 295–308.
- Howe, K. R., & Moses, M. S. (1999). Ethics in educational research. *Review of Research in Education*, 24(2), 21–59.
- Hox, J. J. (2002). The multilevel approach to meta-analysis. *Multilevel Analysis: Techniques and Applications*, 139–156. Retrieved from <http://www.loc.gov/catdir/enhancements/fy0634/2001057760-d.html>
- Hseuh, J., Lowenstein, A. E., Morris, P., & Mattera, S. (2014). Impacts of social-emotional curricula on three-year-olds: Exploratory findings from the Head Start CARES demonstration. Washington DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, United States Department of Health and Human Services.

- 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. Retrieved from <http://www.tandfonline.com/doi/abs/10.1080/10705519909540118>
- Huffman, D. (1997). Effect of explicit problem solving instruction on high school students' problem-solving performance and conceptual understanding of physics. *Journal of Research in Science Teaching*, 34(6), 551–570.
- Hutchings, J., Martin-Forbes, P., Daley, D., & Williams, M. (2013). A randomized controlled trial of the impact of a teacher classroom management program on the classroom behavior of children with and without behavior problems. *Journal of School Psychology*, 51(5), 571–585. Retrieved from <http://dx.doi.org/10.1016/j.jsp.2013.08.001>
- Hyson, M., Copple, C., & Jones, J. (2006). Early childhood development and education. In *Handbook of child psychology: Child psychology in practice* (6th ed., pp. 3–47). Hoboken, NJ, US: Wiley.
- IBM. (2012). SPSS Statistics for Windows. Armonk, New York: IBM Corp.
- Institute of Educational Sciences. (2016). Montessori method: No report available - No studies meeting evidence standards. Retrieved August 24, 2016, from <http://ies.ed.gov/ncee/wwc/Topic.aspx?sid=4>
- Jacob, R., & Parkinson, J. (2015). The potential for school-based interventions that target executive function to improve academic achievement: A review. *Review of Educational Research*, 10(1), 1–41. Retrieved from <http://rer.sagepub.com/cgi/doi/10.3102/0034654314561338>
- Jones, K., Daley, D., Hutchings, J., Bywater, T., & Eames, C. (2007). Efficacy of the Incredible Years basic parent training programme as an early intervention for children with conduct problems and ADHD. *Childcare, Health and Development*, 33(6), 749–756. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/17944785>

- Jones, S., Bub, K., & Raver, C. (2013). Unpacking the black box of the CSRP intervention: The mediating roles of teacher-child relationship quality and self-regulation. *Early Education and Development, 24*, 1043–1064. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/24729666>
- Karreman, A., Van Tuijl, C., & Marcel, A. G. (2006). Parenting and self-regulation in preschoolers: A meta-analysis. *Infant and Child Development, 15*, 561–579.
- Kaufman, S., Reynolds, M., Liu, X., Kaufman, A., & McGrew, K. (2012). Are cognitive g and academic achievement g one and the same g? An exploration on the Woodcock-Johnson and Kaufman tests. *Intelligence, 40*(2), 123–138. <http://doi.org/10.1016/j.intell.2012.01.009>
- Kelley, T. L. (1927). *Interpretation of educational measurements*. New York: World Book Company.
- Kenny, D. (1975). Cross-lagged panel correlation: A test for spuriousness. *Psychological Bulletin, 82*(6), 887–903. Retrieved from <http://content.apa.org/journals/bul/82/6/887>
- Kline, R. (2015). *Principles and practice of structural equation modeling* (4th ed.). New York: Guilford Press.
- Kochanska, G., Coy, K. C., & Murray, K. T. (2001). The development of self-regulation in the first four years of life. *Child Development, 72*(4), 1091–1111.
- Kogan, N., & Carter, A. S. (1996). Mother-infant reengagement following the still-face: The role of maternal emotional availability in infant affect regulation. *Infant Behavior and Development, 19*, 359–370.
- Kopp, C. B. (1982). Antecedents of self-regulation: A developmental perspective. *Developmental Psychology, 18*(2), 199–214.
- Krafft, K. C., & Berk, L. E. (1998). Private speech in two preschools: Significance of open-ended activities and make-believe play for verbal self-regulation. *Early Childhood Research Quarterly, 13*(4), 637–658. Retrieved from

<http://linkinghub.elsevier.com/retrieve/pii/S0885200699800659>

- Kushnir, T., Wellman, H. M., & Gelman, S. A. (2008). The role of preschoolers' social understanding in evaluating the informativeness of causal interventions. *Cognition*, 107(3), 1084–1092.
- Kwon, Y. (2003). A comparative analysis of preschool education in Korea and England. *Comparative Education*, 39(4), 479–491.
- Lahiri, D. B. (1951). A method for sample selection providing unbiased ratio estimates. *International Statistical Institute Bulletin*, 33(2), 133–140.
- Landry, S. H., Miller-Loncar, C. L., Karen, E., & Swank, P. R. (2014). The role of early parenting in children's development of executive processes. *Developmental Neuropsychology*, 21(1), 15–41.
- Leerkes, E. M., Blankson, A. N., & Brien, M. O. (2010). Differential effects of maternal sensitivity to infant distress and non-distress on social-emotional functioning. *Child Development*, 80(3), 762–775.
- Lemov, D. (2010). *Teach like a champion: 49 techniques that put students on the path to college*. San Francisco: Jossey-Base.
- Lengua, L. J. (2003). Associations among emotionality, self-regulation, adjustment problems, and positive adjustment in middle childhood. *Journal of Applied Developmental Psychology*, 24(5), 595–618. Retrieved from <http://linkinghub.elsevier.com/retrieve/pii/S0193397303000960>
- Leong, D., & Bodrova, E. (2011). *Tools of the Mind: Pre-k training manual*. Denver: Metropolitan State College of Denver.
- Liberati, A., Altman, D. G., Gøtzsche, P. C., & Ioannidis, J. P. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *Annals of Internal Medicine*, 151(4), 1–8.

- Liew, J. (2012). Effortful control, executive functions, and education: Bringing self-regulatory and social-emotional competencies to the table. *Child Development Perspectives*, 6(2), 105–111. Retrieved from <http://doi.wiley.com/10.1111/j.1750-8606.2011.00196.x>
- Lillard, A. (2005). *Montessori: The science behind the genius*. New York: Oxford University Press.
- Lillard, A. (2006). The early years: Evaluating Montessori education. *Science*, 313, 1893–1894. <http://doi.org/10.1126/science.1132362>
- Lillard, A. (2012). Preschool children’s development in classic Montessori, supplemented Montessori, and conventional programs. *Journal of School Psychology*, 50(3), 379–401.
- Lillard, A., Lerner, M., Hopkins, E., Dore, R., Smith, E., & Palmquist, C. (2013). The impact of pretend play on children’s development: A review of the evidence. *Psychological Bulletin*, 139(1), 1–34. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/22905949>
- Lin, H. L., Lawrence, F. R., & Gorrell, J. (2003). Kindergarten teachers’ views of children’s readiness for school. *Early Childhood Research Quarterly*, 18(2), 225–237.
- Lipsey, M., & Wilson, D. (2001). *Practical meta-analysis*. London: Sage.
- Littell, J. H., Corcoran, J., & Pillai, V. (2008). *Systematic reviews and meta-analysis*. London: Sage.
- Little, T. (2013). *Longitudinal structural equation modeling*. London: Guilford Press.
- Little, T., Nesselroade, J., & Linderberger, U. (1999). On selecting indicators for multivariate measurement and modeling with latent variables: When “good” indicators are bad and “bad” indicators are good. *Psychological Methods*, 4(2), 192–211.
- Little, T., Slegers, D. W., & Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. *Structural Equation Modeling: A Multidisciplinary Journal*, 13(1), 59–72. http://doi.org/10.1207/s15328007sem1301_3
- Lonigan, C. J., & Phillips, B. M. (2012). Comparing skills-focused and self-regulation focused

preschool curricula: Impacts on academic and self-regulatory skills. In *Society for Research on Educational Effectiveness*. Washington DC.

Luciano, M., Wright, M. J., Smith, G. A., Geffen, L. B., & Martin, N. G. (2001). Genetic covariance among measures of information processing speed, working memory, and IQ. *Behavior Genetics*, 31(6), 581–592.

Luria, A. (1966). *Higher cortical functions in man*. New York: Basic Books.

Luria, A. (1981). *Language and cognition*. (J. V. Wertsch, Ed.). New York: Wiley.

Mackay, P. E. (2013). *The effects of tools of the mind on math and reading scores in kindergarten*. ProQuest Dissertations and Theses. University of Massachusetts - Amherst, Ann Arbor. Retrieved from <http://search.proquest.com/docview/1468983495?accountid=13042>

Malmberg, L.-E., & Flouri, E. (2011). The comparison and interdependence of maternal and paternal influences on young children's behavior and resilience. *Journal of Clinical Child and Adolescent Psychology*, 40(3), 434–444. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/21534054>

Maruyama, P. (1998). *Basics of structural equation modeling*. London: Sage.

McClelland, M., Acock, A. C., Piccinin, A., Rhea, S. A., & Stallings, M. C. (2013). Relations between preschool attention span-persistence and age 25 educational outcomes. *Early Childhood Research Quarterly*, 28(2), 314–324.

McClelland, M., & Cameron, C. E. (2012). Self-regulation in early childhood: Improving conceptual clarity and developing ecologically valid measures. *Child Development Perspectives*, 6(2), 136–142. Retrieved from <http://doi.wiley.com/10.1111/j.1750-8606.2011.00191.x>

McDermott, P. A., Leigh, N. M., & Perry, M. A. (2002). Development and validation of the preschool learning behaviors scale. *Psychology in the Schools*, 39(4), 353–365.

McDonald, C., Cameron, C., Phillips, B. M., Travis, Q. M., Glasney, S., & Morrison, F. J.

- (2011). First graders' literacy and self-regulation gains: The effect of individualizing student instruction. *Journal of School Psychology, 48*(5), 433–455.
- McEwen, B. S., & Stellar, E. (1993). Stress and the individual mechanisms leading to disease. *Archives of Internal Medicine, 153*(18), 2093–2101.
- Meador, D. (2015). Experimental evaluation of the Tools of the Mind Pre-K curriculum fidelity of implementation technical report. Nashville, TN: Peabody Research Institute.
- Meador, D., Turner, K., Lipsey, M., & Farran, D. (2013). Administering measures from the PRI learning-related cognitive self-regulation study. Nashville, TN: Peabody Research Institute.
- Meeks, C. B., & Mauldin, T. (1990). Children's time in structured and unstructured leisure activities. *Lifestyles Family and Economic Issues, 11*(3), 257–281. Retrieved from <http://link.springer.com/10.1007/BF00987003>
- Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equation modeling. *Psychological Methods, 10*(3), 259–284.
- Mihalic, S., Fagan, A. A., Irwin, K., & Ballard, D. (2002). *Prevention replication: Factors for implementation success*. Boulder, Colorado.
- Miller, M., Nevado-Montenegro, A. J., & Hinshaw, S. P. (2012). Childhood executive function continues to predict outcomes in young adult females with and without childhood-diagnosed ADHD. *Journal of Abnormal Child Psychology, 40*(5), 657–68. Retrieved from <http://link.springer.com/article/10.1007/s10802-011-9599-y/fulltext.html>
- Mischel, W. (2014). *The marshmallow test*. London: Transworld.
- Mischel, W., Ebbesen, E., & Zeiss, R. (1972). Cognitive and attentional mechanisms in delay of gratification. *Journal of Personality and Social Psychology, 21*(2), 204–218. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/5010404>

- Miyake, A., & Friedman, N. (2012). The nature and organization of individual differences in executive functions: Four general conclusions. *Current Directions in Psychological Science*, 21(1), 8–14. Retrieved from <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3388901&tool=pmcentrez&rendertype=abstract>
- Miyake, Friedman, N., Emerson, M., Witzki, A., Howerter, A., & Wager, T. (2000). The unity and diversity of executive functions and their contributions to complex “frontal lobe” tasks: A latent variable analysis. *Cognitive Psychology*, 41(1), 49–100. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/10945922>
- Moffitt, T. E., Arseneault, L., & Caspi, A. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences of the United States of America*, 108(7), 2693–2698. Retrieved from <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3041102&tool=pmcentrez&rendertype=abstract>
- Moilanen, K. L., Shaw, D. S., Dishion, T. J., Gardner, F., & Wilson, M. (2009). Predictors of longitudinal growth in inhibitory control in early childhood. *Social Development*, 19(2), 326–347. Retrieved from <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2850123&tool=pmcentrez&rendertype=abstract>
- Morris, P., Matterna, S., Castells, N., Bangser, M., Bierman, K., & Raver, C. (2014). Impact findings from the Head Start CARES: National evaluation of three approaches to improving preschoolers’ social and emotional competence. *Office of Planning, Research and Evaluation Report*. Washington DC.
- Muthén, L. K., & Muthén, B. O. (2012). *Mplus user’s guide*.
- Nampijja, M., Apule, B., Lule, S., Akurut, H., Muhangi, L., Elliott, A. M., & Alcock, K. J. (2010). Adaptation of western measures of cognition for assessing 5-year-old semi-urban Ugandan children. *British Journal of Educational Psychology*, 80(1), 15–30.

- National Institute for Health and Care Excellence. (2013). Antisocial behaviour and conduct disorders in children and young people: Recognition and management. Retrieved January 29, 2015, from <https://www.nice.org.uk/guidance/cg158/chapter/About-this-guideline>
- Neuenschwander, R., Röthlisberger, M., Cimeli, P., & Roebbers, C. M. (2012). How do different aspects of self-regulation predict successful adaptation to school? *Journal of Experimental Child Psychology*, 113(3), 353–371.
<http://doi.org/10.1016/j.jecp.2012.07.004>
- Ning, H. K., & Downing, K. (2010). The reciprocal relationship between motivation and self-regulation: A longitudinal study on academic performance. *Learning and Individual Differences*, 20(6), 682–686. Retrieved from <http://linkinghub.elsevier.com/retrieve/pii/S1041608010001159>
- Nix, R. L., Bierman, K. L., Domitrovich, C. E., & Gill, S. (2013). Promoting children's social-emotional skills in preschool can enhance academic and behavioral functioning in kindergarten: Findings from Head Start REDI. *Early Education and Development*, 24, 37–41. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/24311939>
- Nores, M., Belfield, C. R., Barnett, W. S., & Schweinhart, L. (2005). Updating the economic impacts of the High/Scope Perry preschool program. *Educational Evaluation and Policy Analysis*, 27(3), 245–261. <http://doi.org/10.3102/01623737027003245>
- Nota, L., Soresi, S., & Zimmerman, B. (2004). Self-regulation and academic achievement and resilience: A longitudinal study. *International Journal of Educational Research*, 41(3), 198–215. Retrieved from <http://linkinghub.elsevier.com/retrieve/pii/S0883035505000443>
- Nunes, T., Bryant, P., Evans, D., Bell, D., Gardner, S., Gardner, A., & Carraher, J. (2007). The contribution of logical reasoning to the learning of mathematics in primary school. *British Journal of Developmental Psychology*, 25(1), 147–166. Retrieved from <http://doi.wiley.com/10.1348/026151006X153127>
- Nye, E. (2013). *Effectiveness of the Incredible Years Teacher Classroom Management program on student*

and teacher outcomes: A systematic review and meta-analysis (Master's thesis). University of Oxford.

- O'Shaughnessy, T. E., Lane, K. L., Gresham, F. M., & Beebe-Frankenberger, M. E. (2003). Children placed at risk for learning and behavioral difficulties: Implementing a school-wide system of early identification and intervention. *Remedial and Special Education, 24*(1), 27–35.
- Office of the State Superintendent of Education. (2013). *Reducing out-of-school suspensions and expulsions in District of Columbia public and public charter schools*. Washington DC.
- Osborne, A. F., Butler, N. R., & Morris, A. C. (1984). *The social life of Britain's five year olds: A report of the child health and education Study*. London: Routledge.
- Perry, B., & Szalavitz, M. (2006). *The boy who was raised as a dog and other stories from a child psychiatrist's notebook: What traumatized children can teach us about loss, love, and healing*. New York: Basic Books.
- Phillips, M., Goodell, M., Raczynski, J., & Philyaw Perez, A. (2012). Creating and using index scores in the analysis of school policy implementation and impact. *Journal of School Health, 82*(6), 253–261.
- Piaget, J. (1969). *The psychology of the child*. BOOK, London: Routledge.
- Piotrowski, J. T., Lapierre, M., & Linebarger, D. L. (2013). Investigating correlates of self-regulation in early childhood with a representative sample of English-speaking American families. *Journal of Child and Family Studies, 22*(3), 423–436. Retrieved from <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3602616&tool=pmcentrez&rendertype=abstract>
- Polanin, J. R. (2013). *Addressing the issue of meta-analysis multiplicity in education and psychology (Doctoral thesis)*. Loyola University of Chicago.
- Ponitz, C. C., McClelland, M. M., Matthews, J. S., & Morrison, F. J. (2009). A structured observation of behavioral self-regulation and its contribution to kindergarten outcomes.

Developmental Psychology, 45(3), 605–619. Retrieved from
<http://www.ncbi.nlm.nih.gov/pubmed/19413419>

Portilla, X., Ballard, P. J., Adler, N. E., Boyce, W. T., & Obradović, J. (2014). An integrative view of school functioning: Transactions between self-regulation, school engagement, and teacher-child relationship quality. *Child Development*, 85(5), 1915–1931. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/24916608>

Posner, M., & Rothbart, M. (2000). Developing mechanisms of self-regulation. *Development and Psychopathology*, 12(1), 427–441.

Posner, M., & Rothbart, M. (2007). Educating the human brain. In *American Psychological Association* (p. 420). Washington DC.

Preschool Curriculum Evaluation Research Consortium. (2008). Effects of preschool curriculum programs on school readiness. Washington DC: United States Department of Education Institute of Educational Sciences.

Pring, R. (2012). Importance of philosophy in the conduct of educational research. *Journal of International and Comparative Education*, 1(1), 23–30.

Punch, K. (2009). *Introduction to social research methods in education*. London: Sage.

Punch, K. (2014). *Introduction to social research methods: Quantitative and qualitative approaches* (3rd ed.). London: Sage.

Qi, C. H., Kaiser, P., & Milan, S. (2006). Children's behavior during teacher-directed and child-directed activities in Head Start. *Journal of Early Intervention*, 28(2), 97–110. Retrieved from <http://jei.sagepub.com/cgi/doi/10.1177/105381510602800202>

Ramey, C. T., & Campbell, F. A. (1984). Preventive education for high-risk children: Cognitive consequences of the Carolina Abecedarian Project. *American Journal of Mental Deficiency*, 88, 515–523.

Rathunde, K. Csikszentmihalyi, M. (2005). Middle school students' motivation and quality of

- experience: A comparison of Montessori and traditional school environments. *American Journal of Education*, 111(3), 341–371.
- Raver, C. (2012). Low-income children's self-regulation in the classroom: Scientific inquiry for social change. *American Psychologist*, 11, 681–689.
- Raver, C., Blair, C., & Willoughby, M. (2013). Poverty as a predictor of 4-year-olds' executive function: New perspectives on models of differential susceptibility. *Developmental Psychology*, 49(2), 292–304. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/22563675>
- Raver, C., Jones, S., Li-Grining, C., Metzger, M., Smallwood, K., & Sardin, L. (2008). Improving preschool classroom processes: Preliminary findings from a randomized trial implemented in Head Start settings. *Early Childhood Research Quarterly*, 63(3), 253–255. Retrieved from <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2274905&tool=pmcentrez&rendertype=abstract>
- Raver, C., Jones, S., Li-Grining, C., Zhai, F., Bub, K., & Pressler, E. (2011). CSRP's impact on low-income preschoolers' preacademic skills: Self-regulation as a mediating mechanism. *Child Development*, 82(1), 362–378. Retrieved from <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3682645&tool=pmcentrez&rendertype=abstract>
- Regan, K. S., & Martin, P. J. (2013). Cultivating self-regulation for students with mild disabilities: WHAT'S UP? *Intervention in School and Clinic*, 49(3), 164–173. Retrieved from <http://isc.sagepub.com/cgi/doi/10.1177/1053451213496163>
- Rief, S. (2007). Strategies to improve self-regulation. In *Understanding and managing children's classroom behavior: Creating sustainable, resilient classrooms* (2nd ed., pp. 322–360). Hoboken: Wiley and Sons.
- Rimm-Kaufman, S., Pianta, R., & Cox, M. (2000). Teachers' judgments of problems in the transition to kindergarten. *Early Childhood Research Quarterly*, 15(2), 147–166.

- Rosen, C. (1974). The effects of sociodramatic play on problem-solving behavior among culturally disadvantaged preschool children. *Child Development*, 45(4), 920–927.
- Röthlisberger, M., & Neuenschwander, R. (2011). Improving executive functions in 5- and 6-year-olds: Evaluation of a small group intervention in prekindergarten and kindergarten children. *Infant and Child Development*, 4(1), 1–19.
- Ryan, S. (2004). Message in a model: Teachers' responses to a court-ordered mandate for curriculum reform. *Educational Policy*, 18(5), 661–685.
- Saltz, E., Dixon, D., & Johnson, J. (1977). Training disadvantaged preschoolers on various fantasy activities: Effects on cognitive functioning and impulse control. *Child Development*, 48(2), 367–380.
- Sameroff, A. (1975). Transactional models in early social relations. *Human Development*, 18(3), 65–79.
- Sameroff, A. (Ed.). (2009). *The transactional model of development: How children and contexts shape each other*. Washington DC: American Psychological Association.
- Sanders, M. R. (1999). Triple P - Positive Parenting Program: Towards an empirically validated multilevel parenting and family support strategy for the prevention of behavior and emotional problems in children. *Clinical Child and Family Psychology*, 2(2), 71–90.
- Sanders, M. R., & Mazzucchelli, T. G. (2013). The promotion of self-regulation through parenting interventions. *Clinical Child and Family Psychology*, 16(1), 1–17. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/23397366>
- Sarama, J., & Clements, D. (2003). Building blocks of early childhood mathematics. *Teaching Children Mathematics*, 9(8), 480.
- Scaramella, L. V., & Leve, L. D. (2004). Clarifying parent-child reciprocities during early childhood: The early childhood coercion model. *Clinical Child and Family Psychology Review*, 7(2), 89–107. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/15255174>

- Schickedanz, J., & Dickinson, D. (2005). *Opening the world of learning literacy curriculum*. Iowa: Pearson.
- Schlam, T., Wilson, N., Shoda, Y., & Mischel, W. (2013). Preschoolers' delay of gratification predicts their body mass index 30 years later. *Journal of Pediatrics*, 162(1), 90–93.
- Schmitt, S., McClelland, M., Tominey, S., & Acock, A. (2015). Strengthening school readiness for Head Start children: Evaluation of a self-regulation intervention. *Early Childhood Research Quarterly*, 30, 20–31. Retrieved from <http://dx.doi.org/10.1016/j.ecresq.2014.08.001>
- Schraw, G., Crippen, K. J., & Hartley, K. (2006). Promoting self-regulation in science education: Metacognition as part of a broader perspective on learning. *Research in Science Education*, 36(1–2), 111–139. Retrieved from <http://link.springer.com/10.1007/s11165-005-3917-8>
- Schutt, R. (2011). *Investigating the social world: The process and practice of research* (7th ed.). London: Sage.
- Shayer, M. (2003). Not just Piaget; not just Vygotsky, and certainly not Vygotsky as alternative to Piaget. *Learning and Instruction*, 13(5), 465–485. [http://doi.org/10.1016/S0959-4752\(03\)00092-6](http://doi.org/10.1016/S0959-4752(03)00092-6)
- Shumow, L., Vandell, D. L., & Posner, J. K. (1998). Harsh, firm, and permissive parenting in low-income families: Relations to children's academic achievement and behavioral adjustment. *Journal of Family Issues*, 19(5), 483–507. Retrieved from <http://jfi.sagepub.com/cgi/doi/10.1177/019251398019005001>
- Singer, J., & Willett, J. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford University Press.
- Sitzmann, T., & Ely, K. (2011). A meta-analysis of self-regulated learning in work-related training and educational attainment: What we know and where we need to go. *Psychological Bulletin*, 137(3), 421–442. Retrieved from

<http://www.ncbi.nlm.nih.gov/pubmed/21401218>

- Skibbe, L., Connor, C. M., Morrison, F., & Jewkes, A. (2013). Schooling effects on preschoolers' self-regulation, early literacy, and language growth. *Early Childhood Research Quarterly, 26*(1), 42–49.
- Skibbe, L. E., Phillips, B. M., Day, S. L., Brophy-Herb, H. E., & Connor, C. M. (2012). Children's early literacy growth in relation to classmates' self-regulation. *Journal of Educational Psychology, 104*(3), 541–553. Retrieved from <http://doi.apa.org/getdoi.cfm?doi=10.1037/a0029153>
- Skinner, E. (1986). The origins of young children's perceived control: Mother contingent and sensitive behavior. *International Journal of Behavioral Development, 9*, 359–382.
- Skinner, E., Zimmer-gembeck, M. J., & Connell, J. P. (1998). Individual differences and the development of perceived control. *Monographs of the Society for Research in Child Development, 63*(2), 1–231.
- Smilansky, S. (1968). *The effects of sociodramatic play on disadvantaged preschool children*. New York: Wiley.
- Smith-Donald, R., Raver, C., Hayes, T., & Richardson, B. (2007). Preliminary construct and concurrent validity of the Preschool Self-regulation Assessment (PSRA) for field-based research. *Early Childhood Research Quarterly, 22*(2), 173–187. Retrieved from <http://linkinghub.elsevier.com/retrieve/pii/S0885200607000051>
- Smith, E. (2008). Pitfalls and promises: The use of secondary data analysis in educational research. *British Journal of Educational Studies, 56*(3), 323–339.
- Soares, D. A., Vannest, K. J., & Harrison, J. (2009). Computer aided self-monitoring to increase academic production and reduce self-injurious behavior in a child with autism. *Behavior Interventions, 183*(4), 171–183.
- Spinrad, T., Eisenberg, N., Silva, K., Hofer, C., & Smith, C. (2012). Longitudinal relations among maternal behaviors, effortful control and young children's committed

compliance. *Developmental Psychology*, 48(2), 552–566.

Stayton, D. J., & Ainsworth, M. S. (1973). Individual differences in infant responses to brief, everyday separations as related to other infant and maternal behaviors. *Developmental Psychology*, 9(2), 226.

Stechuk, R. (2008). *Predicting language and early literacy outcomes in low-income children: A structural equation modeling approach (Doctoral thesis)*. George Mason University.

Steenbergen, M., & Jones, B. (2002). Modeling multilevel data structures. *American Journal of Political Science*, 46(1), 218–237.

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., & Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *British Medical Journal*, (338), 2393.

Stright, A. D., Neitzel, C., Sears, K. G., & Hoke-Sinex, L. (2001). Instruction begins in the home: Relations between parental instruction and children's self-regulation in the classroom. *Journal of Educational Psychology*, 93(3), 456–466. Retrieved from <http://doi.apa.org/getdoi.cfm?doi=10.1037/0022-0663.93.3.456>

Strommen, E. A. (1973). Verbal self-regulation in a children's game: Impulsive errors on "Simon Says." *Child Development*, 44(4), 849–853.

Sutton, R. E., Mudrey-Camino, R., & Knight, C. C. (2009). Teachers' emotion regulation and classroom management. *Theory Into Practice*, 48(2), 130–137. Retrieved from <http://www.tandfonline.com/doi/abs/10.1080/00405840902776418>

Swanson, H. L. (1999). Instructional components that predict treatment outcomes for students with learning disabilities: Support for a combined strategy and direct instruction model. *Learning Disabilities Research and Practice*, 14(3), 129–140.

Tabachnick, B., & Fidell, L. (2013). *Using multivariate statistics* (6th ed.). London: Pearson.

Tanner-Smith, E. E., & Tipton, E. (2014). Robust variance estimation with dependent effect

- sizes: Practical considerations including a software tutorial in Stata and SPSS. *Research Synthesis Methods*, 5(1), 13–30. <http://doi.org/10.1002/jrsm.1091>
- Tarullo, A. R., Obradovic, J., & Gunnar, M. R. (2009). Self-control and the developing brain. *Zero to Three*, 29(3), 31–37.
- Teaching Strategies. (2010). *The Creative Curriculum® for preschool*. Washington DC.
- Thompson, A. M., Ruhr, L. R., Maynard, B. R., Pelts, M. D., & Bowen, N. K. (2013). Self-management interventions for reducing challenging behaviors among school-age students: A systematic review. *Campbell Systematic Reviews*, 1–50.
- Tominey, S. L., & McClelland, M. M. (2011). Red light, purple light: Findings from a randomized trial using circle time games to improve behavioral self-regulation in preschool. *Early Education & Development*, 22(3), 489–519.
- Tough, P. (2012). *How children succeed*. Boston: Houghton Mifflin Harcourt.
- Tronick, E. Z. (1989). Emotions and emotional communication in infants. *American Psychologist*, 44(2), 112–119.
- Turner, K. A., Lipsey, M. W., Fuhs, M. W., Vorhaus, E., & Meador, D. N. (2012). Academically relevant measures of executive function: Development and validation of assessments for preschool children. In *Annual Conference of the Society for Research on Educational Effectiveness* (pp. 8–11). Washington DC.
- Turque, B. (2011, November 11). D.C. school reform targets early lessons. *Washington Post*, pp. 1–3. Washington DC.
- United States Department of Education. (2008). *What Works Clearinghouse: Tools of the Mind*. Washington DC.
- United States Department of Education. (2013a). *What Works Clearinghouse: Bright Beginnings*. *What Works Clearinghouse*.

- United States Department of Education. (2013b). *What Works Clearinghouse: The Creative Curriculum for Preschool*. Washington DC.
- Ursache, A., Blair, C., & Raver, C. (2012). The promotion of self-regulation as a means of enhancing school readiness and early achievement in children at risk for school failure. *Child Development Perspectives*, 6(2), 122–128.
- Valentine, J. C., Pigott, T. D., & Rothstein, H. R. (2010). How many studies do you need?: A primer on statistical power for meta-analysis. *Journal of Educational and Behavioral Statistics*, 35(2), 215–247. <http://doi.org/10.3102/1076998609346961>
- van der Veer, R. (2007). *Lev Vygotsky*. (R. Bailey, Ed.). London: Continuum.
- Velayutham, S., Aldridge, J., & Fraser, B. (2011). Development and validation of an instrument to measure students' motivation and self-regulation in science learning. *International Journal of Science Education*, 33(15), 2159–2179. Retrieved from <http://www.tandfonline.com/doi/abs/10.1080/09500693.2010.541529>
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3), 1–48.
- Vorhaus, E., & Meador, D. (2010). Tools of the Mind curriculum implementation fidelity checklist. Nashville, TN: Peabody Research Institute. Retrieved from <https://my.vanderbilt.edu/toolsofthemindevaluation/>
- Vygotsky, L. (1930). Socialisticheskaja peredelka cheloveka (The socialist alteration of man). *Varnitso*, 1–5.
- Vygotsky, L. (1933a). Dinamika umstvennogo razvitiya shkol'nika v svyazi s obucheniem (Dynamics of child development in school). In *Umstvennoe razvitie detey v protsesse obucheniya (Education processes and child development)* (pp. 33–52). Moscow-Leningrad: Uchpedgiz.
- Vygotsky, L. (1933b). Play and its role in the mental development of the child. *Soviet*

Psychology, 6, 5–22.

Vygotsky, L. (1962). *Thought and language*. Cambridge: MIT Press.

Vygotsky, L. (1978). *Mind in society: The development of higher psychological processes*. Cambridge: Harvard University Press.

Vygotsky, L. (1981). The genesis of higher mental functions. In J. V Wertsch (Ed.), *The concept of activity in Soviet Psychology* (pp. 144–188). Armonk, New York: M. E. Sharpe.

Vygotsky, L. (1994). The development of academic concepts in school age children. In R. van der Veer (Ed.), *The Vygotsky Reader* (pp. 355–370). Cambridge, MA: Oxford University Press.

Vygotsky, L. (1997). Development of personality and world view in the child. In R. W. Rieber (Ed.), *Collected works of L. S. Vygotsky: Vol. 4. The history of the development of higher mental functions* (pp. 241–251). New York: Plenum.

Vygotsky, L. (1998a). Development of higher mental functions during the transitional age. In R. W. Rieber (Ed.), *Child psychology, Vol. 5., Collected works of L. S. Vygotsky* (pp. 83–149). New York: Plenum.

Vygotsky, L. (1998b). The problem of age. In R. W. Rieber (Ed.), *Child psychology, Vol. 5., Collected works of L. S. Vygotsky* (pp. 187–205). New York: Plenum.

Vygotsky, L., & Luria, A. (1994). Tool and symbol use in child development. In R. van der Veer (Ed.), *The Vygotsky Reader* (pp. 99–174). Cambridge, MA: Blackwell.

Weaver, G. C. (1998). Strategies in K-12 science instruction to promote conceptual change. *Science Education*, 82(4), 455–472.

Webster-Stratton, C., Reid, M. J., & Hammond, M. (2004). Treating children with early-onset conduct problems: Intervention outcomes for parent, child, and teacher training. *Journal of Clinical Child and Adolescent Psychology*, 33(1), 105–124.

- Webster-Stratton, C., Reid, M., & Stoolmiller, M. (2009). Preventing conduct problems and improving school readiness: Evaluation of the Incredible Years teacher and child training programs in high-risk schools. *Journal of Child Psychology and Psychiatry*, 49(5), 471–488.
- Weiland, C., & Yoshikawa, H. (2013). Impacts of a prekindergarten program on children's mathematics, language, literacy, executive function, and emotional skills. *Child Development*, 84(6), 2112–2130. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/23534487>
- Weisberg, D. S., Hirsh-Pasek, K., & Golinkoff, R. M. (2013a). Embracing complexity: Rethinking the relation between play and learning. *Psychological Bulletin*, 139(1), 35–39. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/23294088>
- Weisberg, D. S., Hirsh-Pasek, K., & Golinkoff, R. M. (2013b). Guided play: Where curricular goals meet a playful pedagogy. *Mind, Brain, and Education*, 7(2), 104–112. Retrieved from <http://doi.wiley.com/10.1111/mbe.12015>
- Whiteside, S. P., & Lynam, D. R. (2001). The five factor model and impulsivity: Using a structural model of personality to understand impulsivity. *Personality and Individual Differences*, 30(4), 669–689. Retrieved from <http://linkinghub.elsevier.com/retrieve/pii/S0191886900000647>
- Wiebe, S., Espy, K., & Charak, D. (2008). Using confirmatory factor analysis to understand executive control in preschool children. *Developmental Psychology*, 44(2), 575–587. <http://doi.org/10.1037/0012-1649.44.2.575>
- Wiebe, S., Sheffield, T., & Nelson, J. (2011). The structure of executive function in 3-year-old children. *Journal of Experimental Child Psychology*, 108(3), 436–452. <http://doi.org/10.1016/j.jecp.2010.08.008>.The
- Williamson, P. R., Gamble, C., Altman, D. G., & Hutton, J. L. (2005). Outcome selection bias in meta-analysis. *Statistical Methods in Medical Research*, 14(5), 515–524.

- Willoughby, M., Holochwest, S., Blanton, Z., & Blair, C. (2014). Executive functions: Formative versus reflective measurement. *Measurement, 12*(3), 69–95.
- Wilson, D. B., Olaghere, A., & Gill, C. (2016). Juvenile curfew effects on criminal behavior and victimization: A Campbell Collaboration systematic review. *Journal of Experimental Criminology, 12*(2), 167–186. <http://doi.org/10.1007/s11292-016-9258-y>
- Wilson, S., & Farran, D. (2012). Experimental evaluation of the Tools of the Mind preschool curriculum. In *Society for Research on Educational Effectiveness* (Vol. 74, pp. 1–9). Washington DC.
- Wilson, S. J., & Lipsey, M. W. (2007). School-based interventions for aggressive and disruptive behavior: Update of a meta-analysis. *American Journal of Preventive Medicine, 33*(2), 1–29. <http://doi.org/10.1016/j.amepre.2007.04.011>
- Winsler, A., Manfra, L., & Diaz, R. (2007). “Should I let them talk?”: Private speech and task performance among preschool children with and without behavior problems. *Early Childhood Research Quarterly, 22*(2), 215–231. Retrieved from <http://linkinghub.elsevier.com/retrieve/pii/S088520060700004X>
- Wolfe, R. N., & Johnson, S. D. (1995). Personality as a predictor of college performance. *Educational and Psychological Measurement, 55*(2), 177–185.
- Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry, 17*(2), 89–100.
- Wu, A. D., Liu, Y., Gadermann, A. M., & Zumbo, B. D. (2010). Multiple-indicator multilevel growth model: A solution to multiple methodological challenges in longitudinal studies. *Social Indicators Research, 97*, 123–142. <http://doi.org/10.1007/s11205-009-9496-8>
- Zelazo, P. (2006). The dimensional change card sort (DCCS): A method of assessing executive function in children. *Nature Protocol, 1*, 297–301.
- Zelazo, P., Anderson, J. E., Richler, J., Wallner-Allen, K., Beaumont, J. L., & Weintraub, S. (2013). NIH toolbox cognition battery (CB): Measuring executive function and

attention. *Monographs of the Society for Research in Child Development*, 78(4), 16–33.

Zelazo, P., & Carlson, S. (2012). Hot and cool executive function in childhood and adolescence: Development and plasticity. *Child Development Perspectives*, 6(4), 354–360.
Retrieved from <http://doi.wiley.com/10.1111/j.1750-8606.2012.00246.x>

Zimmerman, B. (2008). Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects. *American Educational Research Journal*, 45(1), 166–183.

Appendix A: Meta-analysis coding form

Once the seven studies were retrieved from the systematic search, I independently coded each study using the coding form below. The coding form enables systematic data extraction by codifying which study characteristics should be noted as well as how they should be coded in the codebook. Thus, the pages below contain the complete coding form, which I adapted from Littell et al. (2008)

Tools meta-analysis coding form

Item	Description	Notes
Section 1: Study Identification		
1	Study ID:	
2	Author(s) and year: e.g., Bodrova & Leong, 2007	
3	Type of report (select one) 1) Journal article 2) Book/book chapter 3) Government report (e.g., federal, state, local) 4) Thesis or dissertation 5) Conference proceedings 6) Unpublished 7) Other (specify)	
Section 2: Study Context		
1	Country in which the study was conducted: 1) USA 2) Canada 3) Other country (specify)	

2	Regional location of the research site: 1) Suburban 2) Urban 3) Rural 4) Mixed 5) Cannot tell	
Section 3: Sample Description		
1	Number of students (for treatment group, comparison group, and total)	
2	Child gender (0 = female, 1 = male)	
3	Child age (0 = pre-kindergarten, 1 = kindergarten)	
4	Special education status (0 = no, 1 = yes)	
5	Ethnicity information (as described in the study)	
6	Socio-economic status (as described in the study)	
7	English language learners (as described in the study)	
8	Participant attrition rate (treatment group, comparison group, or two groups combined)	
9	Reason for attrition (as described in the study)	
Section 4: Description of intervention and comparison condition		

1	Comparison condition (as described in the study)	
2	Were efforts made to monitor and measure fidelity of implementation? 1) Yes (how) • Observations • Interviews of participants • Surveys of participants • Participant logs • Administrative records • Checklists • Other 2) No	
3	Duration/frequency of Tools implementation (as described in the study)	
Section 5: Research Design		
1	Research design type: 1) Experimental design (included randomized controlled trials or cluster-randomized trials) 2) Quasi-experimental design– Regression discontinuity, differences-in-differences, instrumental variables 3) Quasi-experimental design– two groups, pre-and post-test design 4) Quasi-experimental design– two groups, post-test only (no pre-test) 5) Longitudinal study–outcomes were measured at least twice after intervention	

2	Unit of assignment to conditions: 1) Individual 2) Group/cluster/sites (specify)	
3	Unit of analysis: 1) Individual 2) Group/cluster/sites (specify)	
4	Method of assignment to conditions: 1) Completely random 2) Random after matching, stratification, blocking, etc. 3) Quasi-random-assigned by some naturally existing situations 4) Nonrandom, but matched or statistically controlled on major characteristics or pretest measures	
5	If matching was used, how were the groups matched? (select all that apply) 1) Matched on pretest measures 2) Matched on demographics or other major features 3) Propensity score matching	
6	Were the participants (i.e., teachers and children) blinded to their conditions? 1) Yes 2) No	

7	Was the data collector blind to the group assignment? 1) Yes 2) No	
8	Results of statistical comparisons of pre-intervention group differences 1) No statistically significant differences 2) Statistically significant differences 3) No comparisons were made	
9	Upon what kind of the statistical analyses were the major findings of the original study based? 1) Descriptive analysis 2) <i>t</i> -tests 3) ANOVA/MANOVA 4) ANCOVA/MANCOVA 5) Regression/multiple regression 6) Factor analysis 7) Path analysis 8) Multilevel modeling 9) Structural equation modeling (SEM) 10) Other (specify)	
Section 6: Outcome Measures		
1	Outcome measures (select all that apply) 1) Achievement/learning outcome measures (e.g., standardized test scores, course grades)	

	2) Performance-based executive function tests (e.g., inhibitory control, working memory, cognitive flexibility) 3) Rating scales, survey, questionnaire, and checklist 4) Behavioral observation	
2	Source of outcome data: 1) Child 2) Parent report 3) Teacher report/caregiver report 4) Other	
4	Were the reliability and validity of the outcome measures reported in the study? 1) Yes (specify) 2) No	
5	When did the post-test measure(s) take place? 1) Immediately following the intervention 2) Follow-up/delayed (specify)	
6	Quantitative information on outcomes of interests (e.g., means, standard deviations, <i>t</i> -values) (Note: all related outcomes will be extracted from the study and will be recorded in an Excel file for effect size calculations)	
7	Effect size calculation	

	(e.g., Hedges' g , odd ratio, page number where the related original outcome data located, corresponding to each calculated effect sizes)	
Section 7: Coding Information		
1	Coder	
2	Coding time: How much time (minutes) does it take to complete the coding?	
3	Date of coding	
4	Coding agreement rate with another independent coder (%)	
5	Areas/reasons of coding discrepancies (specify)	
6	How coding discrepancies were resolved (specify)	

Appendix B: Sample syntax for R and SPSS

As described in Chapter Eight, the meta-analysis was conducted through a multilevel framework as well as a robust variance estimation (RVE) framework. The multilevel meta-analysis was conducted using the R package called metafor (Viechtbauer, 2010). The syntax is depicted below.

```
> library(metafor)
> obj <- read.csv("/Users/abaro2/Documents/DPhil/DPhil
Writing/Meta-analysis/MLM_ES_table_Obj.csv")
> View(obj)
> MLM <-
rma.mv(yi=effectsize,V=var,data=obj,random=list(~1|esid,~1|stud
id))
> summary(MLM)
```

The full output from R Studio for the multilevel meta-analysis model with the assessor-reported self-regulation data is as follows:

Multivariate Meta-Analysis Model (k = 12; method: REML)

logLik	Deviance	AIC	BIC	AICc
13.7438	-27.4876	-21.4876	-20.2939	-18.0590

Variance Components:

estim	sqrt	nlvls	fixed	factor
sigma^2.1	0.0000	0.0000	12	no esid
sigma^2.2	0.0710	0.2664	3	no studyid

Test for Heterogeneity:

Q(df = 11) = 15.8510, p-val = 0.1468

Model Results:

estimate	se	zval	pval	ci.lb	ci.ub
0.1722	0.1607	1.0716	0.2839	-0.1427	0.4871

In the output above, the log-likelihood (loglik), Deviance, AIC, BIC, and AICc are all fit indices to compare the relative appropriateness of various nested model specifications. This meta-analysis does not compare any nested models with various predictors because moderation analysis through meta-regression was not conducted; thus, the model fit indices do not provide useful information for these meta-analytic results.

Below the model fit indices section in the output, the variance components row indicates the amount of variance observed at different levels of analysis. Those components are denoted as σ^2 (σ^2) in R Studio, whereas some other texts and software packages refer to those components as tau-squared (τ^2) values. In R Studio, the σ^2 (σ^2) at level one indicates the amount of shared variance among effect sizes from all studies, whereas σ^2 at level two indicates the amount of shared variance among effect sizes from the same study.

Thus, as we would expect, there are small but observable values of shared variation among effect sizes from the same studies (σ^2 at level two) because those effect sizes are based on information from the same participants. However, they are capturing different pieces of information about the participants, so we would not expect their shared variation to be extremely high. In the output above, 7.1% is the amount of shared variation among effect sizes from the same cluster (i.e., study).

By contrast, we would not expect any additional shared variation among all effect sizes from all studies at level one. Thus, σ^2 at level one is, as expected, zero. Once again, this number quantifies the amount of shared variation across all effect sizes that is observed above and beyond the prediction of sampling error. Since there is no reason to expect shared variation among the twelve effect sizes in the analysis above and beyond that among effect sizes clustered within the same study, the σ^2 value is 0.

Beneath the variance components analysis, we observe the Q-statistic value, whose equation and interpretation are described in the ‘Heterogeneity analysis’ section of Chapter Eight. The Q-statistic in the output above is relatively small and not statistically significant, which indicates that different studies did not reach significantly different conclusions regarding Tools’ effectiveness on children’s assessor-based self-regulation scores.

Finally, in the results above, the final row entitled ‘Model Results’ indicates the composite effect size, its standard error, the Z- and p-values, and the 95% confidence interval. The output indicates a small to moderate effect size ($g = .17$) with a confidence interval that crosses zero, which indicates a lack of statistical significance (the p-value is .28).

Once the multilevel meta-analysis had been conducted in R Studio, I assessed the robustness of the findings using the robust variance estimation (RVE) method. SPSS (IBM, 2012) can run RVE through a software macro with the following syntax.

```

DEFINE ROBUST ( STUDYID !CHAREND ("")
               / EFFSIZE !CHAREND ("")
               / VAREFFS !CHAREND ("")
               / RHO    !CHAREND ("") !DEFAULT ("")
               / DESIGN !CHAREND ("") !DEFAULT ("")
               / WEIGHTS !CHAREND ("") !DEFAULT ("")
               / RESID  !CHAREND ("") !DEFAULT ("")
               / HWEIGHT !CHAREND ("") !DEFAULT ("")
               / PRINT  !CHAREND ("") !DEFAULT (DEF) ).
PRESERVE.
SET MPRINT OFF .
SET PRINTBACK OFF.

ROBUST STUDYID = studyid / EFFSIZE = es / VAREFFS = var /
RHO = .8 .

```

The output from the RVE syntax provides nearly identical information to that of the R Studio output as seen below.

Parameter Estimates and Robust Standard Errors					
	Coef	SE	T	Pr > T	95% Conf Interval
INTERCEP	.060563	.019468	3.110938	.035839	.006512
	.114614				

```

N Level 1
15

N Level 2
5

Average Level 1 N
3.00

T-Test DF
4

Tau-squared estimate
.000000

Assumed Rho
.80

Weighted Residual Sum of Squares Qe
3.234

```

Specifically, the output shows the parameter estimates in the first row (instead of the last row in R Studio). The estimates include the composite effect size estimate, its standard error, the t- and p-values, and the 95% confidence interval. The tau-squared (τ^2) value of zero in the RVE is the same as the σ^2 value of zero at level one in the R Studio multilevel meta-analysis. The value of zero indicates the proportion of shared variance across effect sizes above and beyond that which would be expected by sampling error.

Instead, the shared variation exists among effect sizes nested within the same cluster, as was observed in the non-zero values of σ^2 from the level two in the multilevel meta-analysis. In RVE, however, there is only one level of analysis (i.e., it is not a multilevel model), so the RVE specification does not estimate a σ^2 value at level two. Instead, the researcher must explicitly specify a rho (ρ) value, which estimates the inter-correlation among effect sizes nested within the same cluster (i.e., study).

Thus, I specified a rho value of .80 in the last line of the syntax, which signifies a very high dependency among effect sizes from the same study. This high inter-correlation value

imposes a conservative estimation process on the analysis, which, in turn, reduces the likelihood of a Type I error. The rho value of .80 is the recommendation of the SPSS coders who created the RVE macro (Tanner-Smith & Tipton, 2014) and has been also recommended in other RVE literature (Hedges, Tipton, & Johnson, 2010).

Nonetheless, I also performed a robustness check with low ($\rho = .20$), medium ($\rho = .50$), and high ($\rho = .80$) rho values of assumed inter-correlation. Neither the beta coefficients nor the significance values changed across models, so the results from the RVE, which were themselves a robustness check of the multilevel meta-analysis, can be said to be robust to different estimates of inter-correlations among effect sizes nested in the same study.

Appendix C: DREC approval email

Dear Alex

How to develop big egos in small children: Educational practices predictive of self-regulation development in early childhood

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.

If your research involves participants whose ability to give free and informed consent is in question (this includes those under 18 and vulnerable adults), then it is advisable to read the following NSPCC professional reporting requirements for cases of suspected abuse

http://www.nspcc.org.uk/Inform/research/questions/reporting_child_abuse_wda74908.html

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 research study.

Yours sincerely

Nigel

Dr Nigel Fancourt

Departmental lecturer

Department of Education
15 Norham Gardens, Oxford OX2 6PY
01865 274259
www.education.ox.ac.uk

Appendix D: Systematic search terms and results

This appendix contains the search terms and results for the systematic review outlined in Section III of the dissertation. The systematic search involved 11 electronic databases. The relevant search term and search results for each of the 11 databases are shared alphabetically in the sections below.

The systematic search was conducted on October 21, 2016 beginning at precisely 9 am British Standard Time in the United Kingdom. Whereas the search results may have changed since then (i.e., newer papers could have been added or removed from databases), the results presented below represent exactly what was recovered on the day of the search.

Applied Social Sciences Index and Abstracts (ProQuest)

Search term: AB(“Tools of the Mind” OR TI(“Tools of the Mind”))

Results: 2 hits

CENTRAL (Cochrane Library)

Search term: “Tools of the Mind”

Results: 0 hits

Embase (Ovid: 1947 to October week 2 2016)

Search term: “Tools of the Mind”

Results: 0 hits

ERIC (ProQuest)

Search term: AB(“Tools of the Mind” OR TI(“Tools of the Mind”))

Results: 22 hits

LILACS (<http://lilacs.bvsalud.org/en/>)

Search term: “Tools of the Mind”

Results: 0 hits

MEDLINE (Ovid: 1946 to 20 October 2016)

Search term: “Tools of the Mind”

Results: 4 hits

OpenGrey (www.opengrey.eu/)

Search term: “Tools of the Mind”

Results: 0 hits

PsycINFO (Ovid: 1967 to October week 2 2016)

Search term: “Tools of the Mind”

Results: 22 hits

ProQuest Dissertations and Theses (ProQuest)

Search term: AB(“Tools of the Mind” OR TI(“Tools of the Mind”))

Results: 7 hits

Social Sciences Citation Index (ProQuest)

Search term: AB(“Tools of the Mind” OR TI(“Tools of the Mind”))

Results: 6 hits

Sociological Abstracts (ProQuest)

Search term: AB(“Tools of the Mind” OR TI(“Tools of the Mind”))

Results: 0 hits

Appendix E: Risk of bias evidence tables

Each study included in the systematic review received risk of bias ratings as described in section 9.6. Those ratings were mostly based on textual evidence from the original study. In instances where the information could not be gleaned from the study, I followed up with the authors by email to retrieve the necessary information. If the authors could not be reached or refused to provide information, then the study received an “Unclear risk” designation for that risk of bias dimension. The seven risk of bias tables for the seven included studies are presented alphabetically below.

Barnett et al., 2008

Bias	Author's judgment	Support for judgment
Random sequence generation	Low-risk	"The randomization was by computer generated sequence" (author email).
Allocation concealment	Low-risk	"The researcher who conducted the assignments was the project coordinator who was responsible for organizing data collection; that person was not involved in other aspects of the research including specification of hypotheses, design or analysis" (author email).
Blinding of participants and personnel	Unclear risk	It is not possible to blind teachers and students to their condition assignments.
Blinding of outcome assessment	Unclear risk	It was not possible to blind teachers to their students' curricular assignment when completing the Social Skills Rating Scale (SSRS).

Incomplete outcome data	Low risk	"It was not possible to conduct extensive analyses of attrition, because most attrition in this study was due to lack of active consent from parents prior to any data collection. However, we do know gender, ethnicity, and home language for most of the original sample children. Thus, it was possible to test for differences between those whose parents agreed to participate and those whose parents declined or did not respond. Analysis of Variance revealed no statistically significant main effects of attrition or interactions between attrition and treatment (curriculum assignment)" (Barnett et al., 2008, p. 303).
Selective reporting	Low risk	All outcome scores from the measures described in the methods section are available in the tables

Blair & Raver, 2014

Bias	Author's judgment	Support for judgment
Random sequence generation	Low risk	"The randomization was computer generated" (author email).
Allocation concealment	Low risk	"The randomization was conducted independently by someone not associated with study" (author email).
Blinding of participants and personnel	Unclear risk	It is not possible to blind teachers and students to their condition assignments.
Blinding of outcome assessment	Unclear risk	"The outcome assessors may have been aware of the group assignment of the school. I can't say for sure, one way or the other, but I expect that some of them were" (author email).

Incomplete outcome data	Low risk	"I did [assess differences between attrited and non-attrited students] and differences were minimal" (author email).
Selective reporting	Low risk	All outcome scores from the measures described in the methods section are available in the tables

Clements & Sarama, 2014

Bias	Author's judgment	Support for judgment
Random sequence generation	Low-risk	"Schools/centers were randomly assigned to the three conditions three at a time starting at a randomly chosen point in the sorted list and then moving to the top of the list. This is an application of the systematic circular sampling scheme (Lahiri, 1951), which was utilized to ensure three experimental groups that are balanced geographically and in terms of the length of the Pre-K program and key background characteristics of the schools/centers." (Clements & Sarama, 2014, p. 8).
Allocation concealment	Unclear risk	Not reported
Blinding of participants and personnel	Unclear risk	It is not possible to blind teachers and students to their condition assignments.

Blinding of outcome assessment	Unclear risk	"The stories that children retell during the RBS assessment are transcribed and scored on a series of dimensions by trained coders naïve to the group assignment of the child" (p. 23). However, the assessments were conducted in the classrooms, which would have several environmental cues (e.g., mediators, play center materials, etc.) as to whether or not it was a Tools classroom.
Incomplete outcome data	High risk	"Table 5 presents the corresponding results. The first column in this table shows that the size of the analytic samples for these analyses is roughly ten percent smaller than those of the Spring 2011 measures, where the reduction in the sample size is mostly due to children's mobility between the two time points" (p. 32-33). Thus, the authors note substantial attrition but do not analyze it, hence the high risk rating.
Selective reporting	Low-risk	All outcome scores from the measures described in the methods section are available in the tables

Diamond et al., 2007

Bias	Author's judgment	Support for judgment
Random sequence generation	Low-risk	Computer randomization (Author email)
Allocation concealment	Unclear risk	Not reported
Blinding of participants and personnel	Unclear risk	It is not possible to blind teachers and students to their condition assignments.

Blinding of outcome assessment	High risk	"The intent was for testers to be blinded to condition, but the testers said they could tell which children were Tools because when it came to the most difficult test conditions, control children tended to give up, but Tools children kept saying, "I know I can do this"" (author email).
Incomplete outcome data	High risk	"One entire school left the study. We did not do post-intervention assessments there" (author email).
Selective reporting	Low risk	All outcome scores from the measures described in the methods section are available in the appendices

Farran & Wilson, 2014

Bias	Author's judgment	Support for judgment
Random sequence generation	Low risk	"We used a computer random number generator (in excel) to perform the randomization" (author email).
Allocation concealment	Low risk	"All schools were recruited prior to assignment and all schools were randomized in a single randomization using the procedure described above. So, because knowledge of one assignment could not have affected recruitment or future assignments, allocation was effectively concealed – schools and the researchers were unaware of assignments or upcoming assignments because it was all done at once" (author email).
Blinding of participants and personnel	Unclear risk	It is not possible to blind teachers and students to their condition assignment.

Blinding of outcome assessment	Unclear risk	"I think that for the most part assessors were blind to condition when completing the SAR. We did have some assessors who had also been observers in the pre-K classrooms, so if the assessor went to assess some children in the same classroom that they had observed previously, it would have been obvious to the assessor that those children were in a Tools or Control classroom. Also, I guess, the assessor could have noticed Tools materials, centers, etc. in the classroom when they went to pull the child for the assessment. But the assessment materials (roster of children's names and filemaker system for collecting assessment data) did not indicate if the classroom was Tools or Control. Also, this would have only occurred during the pre-K assessments, in kindergarten and first grade the children had moved into different classrooms and so assessors wouldn't have known if they were in Tools or Control during their pre-K year" (author email).
Incomplete outcome data	Low-risk	"Attrition during the study was minimal. No teachers dropped out during the test year. Attrition of students over the course of the study was low and similar across Tools and comparison classrooms" (Farran & Wilson, 2014, p. 11); "There were no statistically significant differences in attrition by condition" (p. 11).
Selective reporting	Low-risk	All outcome scores from the measures described in the methods section are available in the appendices

Lonigan & Phillips, 2012

Bias	Author's judgment	Support for judgment
Random sequence generation	Unclear risk	Not reported
Allocation concealment	Unclear risk	Not reported
Blinding of participants and personnel	Low risk	It is not possible to blind teachers and students to their condition assignment.
Blinding of outcome assessment	Unclear risk	"Children's classroom teachers completed the Behavioral Rating Inventory of Executive Function - Preschool version" (3). The task-based executive function test appears to have been conducted by blind assessors, though it is unclear.
Incomplete outcome data	Unclear risk	Not reported
Selective reporting	Low risk	All outcome scores from the measures described in the methods section are available in the tables

Morris et al., 2014

Bias	Author's judgment	Support for judgment
Random sequence generation	Unclear risk	Not reported
Allocation concealment	Unclear risk	Not reported
Blinding of participants and personnel	Unclear risk	It is not possible to blind teachers and students to their condition assignments.
Blinding of outcome assessment	Unclear risk	Not reported
Incomplete outcome data	Unclear risk	Not reported

Selective reporting	Low risk	All outcome scores from the measures described in the methods section are available in the tables.

Appendix F: E-mail from PRI to confirm dataset transfer

The following is the e-mail text from the research manager at Vanderbilt University's Peabody Research Institute. The e-mail confirms the dataset transfer and requests that I not share the datasets with anyone else. The file descriptions and passwords have been blacked out to ensure confidentiality.

Hi Alex,

Here are the links to access the de-identified fidelity dataset and a limited version of our narrative dataset (included so you have access to length and proportion of time for the Tools timeblocks). Now for the disclaimers: These datasets cannot be shared with anyone and you cannot publish any information from these datasets without permission from PRI. Please let us know if you have any questions and perhaps we'll try to conference again next month once you have had some time to dig in to them. Thanks and have a great weekend. Enjoy!

Deanna

Description: [REDACTED]

Password: [REDACTED]

To download the file open the following link in a browser

<https://vshare.vanderbilt.edu/www/?a=d&i=OjNsewyRJ6>

Description: [REDACTED]

Password: [REDACTED]

To download the file open the following link in a browser

<https://vshare.vanderbilt.edu/www/?a=d&i=TMWAd6kt2i>

Appendix G: Exploratory factor analysis results

This appendix contains the exploratory factor analysis (EFA) results for the three self-regulation outcomes: the executive function indicators, the teacher-reported CFBRs indicators, and the researcher-reported SAR indicators. The first section below contains the EFA results for the five executive function indicators at each of the four time points.

Outcome #1: Executive function indicators

The executive function EFA results were obtained using the following sample SPSS syntax for each set of indicators for each time point:

```
FACTOR /VARIABLES  
Indicator_1 Indicator_2 Indicator_3 Indicator_4 Indicator_5  
/PRINT INITIAL KMO REPR EXTRACTION ROTATION  
/FORMAT BLANK(.25)  
/CRITERIA MINEIGEN(1) ITERATE(25)  
/EXTRACTION ML  
/CRITERIA ITERATE(25)  
/ROTATION VARIMAX .
```

The result output below contains two parts for each test: Firstly, I present the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy and Bartlett's test of sphericity. For each test, the results should be KMO values over .6 and Bartlett's p-value < .05, respectively (Tabachnick & Fidell, 2013). Secondly, I present an eigenvalue table. The results below indicate the emergence, at each time point, of only one factor, which is evidenced by only one eigenvalue above 1 in each table (Tabachnick & Fidell, 2013).

Time point one for executive function

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.703
Bartlett's Test of Sphericity	Approx. Chi-Square	735.147
	df	10
	Sig.	0

Factor	Initial Eigenvalues	Percent	
	Total	% of Variance	Cumulative %
1	2.075	41.498	41.498
2	0.91	18.207	59.705
3	0.846	16.915	76.62
4	0.702	14.041	90.661
5	0.467	9.339	100

Time point two for executive function

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.732
Bartlett's Test of Sphericity	Approx. Chi-Square	1052.139
	df	10
	Sig.	0

Factor	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	2.306	46.117	46.117
2	0.905	18.108	64.225
3	0.764	15.279	79.504
4	0.631	12.625	92.129
5	0.394	7.871	100

Time point three for executive function

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.767
Bartlett's Test of Sphericity	Approx. Chi-Square	988.021

	df	10
	Sig.	0

Factor	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	2.336	46.711	46.711
2	0.874	17.485	64.197
3	0.697	13.937	78.134
4	0.612	12.231	90.365
5	0.482	9.635	100

Time point four for executive function

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.76
Bartlett's Test of Sphericity	Approx. Chi-Square	947.87
	df	10
	Sig.	0

Factor	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	2.295	45.904	45.904
2	0.84	16.807	62.711
3	0.799	15.988	78.699
4	0.576	11.527	90.226
5	0.489	9.774	100

Outcome #2: Teacher-reported self-regulation (CFBRs)

Time point one for CFBRs

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.			0.896
Bartlett's Test of Sphericity	Approx. Chi-Square		3941.312
	df		10
	Sig.		0

Factor	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	3.75	75.005	75.005
2	0.43	8.601	83.606
3	0.337	6.749	90.355
4	0.27	5.404	95.759
5	0.212	4.241	100

Time point two for CFBRs

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.			0.901
Bartlett's Test of Sphericity	Approx. Chi-Square		4342.137
	df		10
	Sig.		0

Factor	Initial Eigenvalues		
	Total	% of Variance	Cumulative %

1	3.87	77.393	77.393
2	0.382	7.633	85.027
3	0.301	6.016	91.043
4	0.253	5.051	96.093
5	0.195	3.907	100

Time point three for CFBRs

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.895
Bartlett's Test of Sphericity	Approx. Chi-Square	4560.297
	df	10
	Sig.	0

Factor	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	3.877	77.549	77.549
2	0.443	8.866	86.415
3	0.293	5.854	92.269
4	0.231	4.615	96.884
5	0.156	3.116	100

Time point four for CFBRs

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.895
Bartlett's Test of Sphericity	Approx. Chi-Square	4574.632
	df	10
	Sig.	0

Factor	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	3.892	77.839	77.839
2	0.403	8.057	85.896
3	0.333	6.656	92.552
4	0.207	4.144	96.696
5	0.165	3.304	100

Outcome #3: Researcher-reported self-regulation (SAR)

Time point one for SAR

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.903
Bartlett's Test of Sphericity	Approx. Chi-Square	5505.835
	df	10
	Sig.	0

Factor	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	4.107	82.149	82.149
2	0.304	6.078	88.227
3	0.293	5.856	94.083
4	0.167	3.332	97.415
5	0.129	2.585	100

Time point two for SAR

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.			0.893
Bartlett's Test of Sphericity	Approx. Chi-Square		4167.36
	df		10
	Sig.		0

Factor	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	3.803	76.065	76.065
2	0.399	7.983	84.047
3	0.367	7.335	91.383
4	0.238	4.76	96.142
5	0.193	3.858	100

Time point three for SAR

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.			0.903
Bartlett's Test of Sphericity	Approx. Chi-Square		4907.63
	df		10
	Sig.		0

Factor	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	4.017	80.35	80.35
2	0.3	5.999	86.348

3	0.292	5.832	92.18
4	0.213	4.255	96.435
5	0.178	3.565	100

Time point four for SAR

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.			0.893
Bartlett's Test of Sphericity	Approx. Chi-Square	3907.492	
	df	10	
	Sig.	0	

Factor	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	3.726	74.519	74.519
2	0.435	8.709	83.228
3	0.381	7.611	90.839
4	0.25	4.994	95.833
5	0.208	4.167	100

Appendix H: Model fit indices used in the present study

The goal of statistical modeling is to specify a model that fits the data (Maruyama, 1998). In order to compare the relative fits of different models, various goodness-of-fit (GOF) indices can be used. The indices reported in this analysis include the chi-squared statistic (with its associated degrees of freedom and significance value), the Comparative Fit Index (CFI), the Tucker-Lewis Index (TLI), and the Root Mean Square Error of Approximation (RMSEA).

Chi-squared (χ^2) statistic

The chi-squared statistic is the most commonly used GOF index for SEM (Blunch, 2008). The chi-squared statistic has an associated p-value, which is typically compared against a threshold of $p < .05$ (Field, 2009). Given the null hypothesis that the specified model fits the data, a non-significant p-value (i.e., $p > .05$) thus indicates model fit.

The chi-squared statistic's critical flaw is its susceptibility to influence based on sample size. For large sample sizes, chi-squared statistics often yield significant p-values and thus falsely reject a model's fit (Kenny, 2015), which could represent a Type I error (i.e., the incorrect finding that a model is significantly different from the data and does not fit). Despite this issue, the chi-squared statistic remains a useful GOF index that is used in the computation of many alternative model fit indices (Bryne, 2009), which will now be discussed.

Comparative Fit Index (CFI)

The CFI improves upon the chi-squared statistic by being much less sensitive to sample size (Bryne, 2009). The CFI is normed so that all possible values range between 0 and 1. Strong model fit is indicated by CFI values approaching 1, with values over .9 being mostly acceptable and values over .95 being excellent (Blunch, 2008). The CFI measures the improvement of the specified model over an unconstrained model in which all the variables are assumed to be uncorrelated (Bryne, 2009).

Tucker Lewis Index (TLI)

The TLI, like the CFI, measures the incremental improvement in model fit of the specified model over a less restricted baseline model. The TLI differs from the CFI in two ways. First, the TLI is a non-normed index, which means that its values can fall outside the range of 0 to 1. The second difference between CFI and TLI is that the latter includes a penalty function for complex models. That is, if two models fit the data equally well, but one is more parsimonious than the other, then the simpler model will have a TLI closer to the ideal value of 1 (Bryne, 2012).

Thus, the penalty function guards against the acceptance of models that are too complex for real-world application. Although the penalty function is a desirable quality for a fit index, the CFI and TLI are often highly correlated, and SEM researchers recommend using only one or the other (Kenny, 2015). Thus, this dissertation will report only the CFI, which is more commonly reported (Kenny, 2015). In order to incorporate the penalty function for complex models, this dissertation will also report the Root Mean Square Error of Approximation (RMSEA), which is described below.

Root Mean Square Error of Approximation (RMSEA)

RMSEA is technically a “badness-of-fit index” (Kline, 2015) where a value of 0 indicates optimal model fit (i.e., the model perfectly describes the variance-covariance matrix). RMSEA uses unknown but systematically identified parameters to estimate how well the model would fit the population covariance matrix. The RMSEA, like the TLI, uses a model’s degrees of freedom to become “parsimony-adjusted” (Kline, 2015).

This signifies that simpler models will, *ceteris paribus*, have RMSEA values closer to 0 than their more complex counterparts. Although .05 is often designated as its threshold value, other methodologists have designated RMSEA values < .06 (Browne & Cudeck, 1992) and values < .08 (Hu & Bentler, 1999) as acceptable thresholds for model fit.

Standardized Root Mean Square Residual (SRMR)

The SRMR transforms the unstandardized variance-covariance matrix into a standardized correlation matrix; it is thus defined as the difference between the observed and predicted correlation values. Like the RMSEA, a value of 0 for SRMR indicates perfect model fit, and any value under .08 is considered acceptable (Hu & Bentler, 1999). Unlike the RMSEA, the SRMR does not take the model's degrees of freedom into account, which means that the SRMR has no penalty for complex models. Given the complexity of several models in the present study, it is useful to include both the SRMR and RMSEA to determine whether the results are robust to model complexity (Kline, 2015).

Summary

Although the aforementioned GOF indices each have their own threshold values, these cut-off points are somewhat arbitrary (Tabachnick & Fidell, 2013). As such, these threshold values are not to be regarded as firm cut-off points but rather useful guides to the researcher. If the thresholds are upheld too strictly, then the probability of falsely rejecting a good model will increase (Blunch, 2008). The latent variable structural equation modeling results chapter (Chapter Twelve) will indicate the various model fit statistics for the relevant models run for study two of this dissertation.

Appendix I: Measurement invariance testing

Measurement invariance models assess the longitudinal consistency of measurement for a given construct. The measurement invariance testing involves three progressively strict tests: configural, weak, and strong invariance. Each will be described sequentially below.

Configural invariance freely estimates the factor loadings and intercepts of each indicator; this essentially tests whether the constituent indicators all make up the same latent construct across time (Wu, Li, & Zumbo, 2007). Without establishing configural invariance, it could be the case that the constituent indicators were actually measuring slightly different constructs across different time points (e.g., self-regulation at one time point versus motivation at another).

The next, more strict, invariance test is called weak invariance, which constrains factor loadings to equality across time while freely estimating the intercepts. This model ensures that a one-unit change in the indicator scores corresponds to an equal change in the factor score across time (Wu et al., 2007). That is, an increase of five points on peg tapping should increase a child's latent executive function score by the same amount at each time point. Thus, weak invariance models test equality of slopes, or change in scores, over time. If the factor loadings are not equal across time, then weak invariance does not hold, and latent scores cannot be compared across time points.

Finally, the strong invariance model constrains both factor loadings and intercepts to equality across time. Thus, this model also tests the assumptions of weak invariance, and then it adds an additional assumption. Specifically, the strong model ensures that the same mean score on the observed indicators would result in the same latent variable score across time points (Wu et al., 2007). Thus, the strong invariance model tests both equality of slopes (i.e., factor loadings in the weak invariance model) and equality of intercepts (i.e., the mean scores on the observed indicators). For example, a strong invariance model could demonstrate that a composite score of 25 across the five executive function indicators results in the same latent variable score at time one through four. Without proving intercept and factor loading equality, analyses with latent variables cannot proceed (Kline, 2015).

Appendix J: List of Tools activities and groupings

Literacy

- ❖ Buddy reading
- ❖ Graphics practice
- ❖ Make a rhyme
- ❖ Take away sounds
- ❖ I have who has letters
- ❖ Elkonin boxes
- ❖ Story lab activities

Math

- ❖ Numberline hop
- ❖ Numerals game
- ❖ Attribute game
- ❖ Complete and continue
- ❖ Making collections
- ❖ Math memory
- ❖ Venger drawing
- ❖ I have who has colors
- ❖ I have who has shapes
- ❖ I have who has numbers
- ❖ Remember and replicate
- ❖ Puzzles and manipulatives

Make-believe play

- ❖ Make-believe play planning
- ❖ Make-believe play practice
- ❖ Make-believe play centers

- ❖ Make-believe play cleanup

Attention-focusing activities

- ❖ Attention focus
- ❖ Pattern movement game
- ❖ Pretend transition
- ❖ Number follow the leader
- ❖ Mouse trap
- ❖ What are you doing, Mr. Wolf?
- ❖ Two-step freeze
- ❖ Freeze on number
- ❖ Partner freeze
- ❖ Freeze game

Science

- ❖ Science eyes

Introduction

- ❖ Message of the day
- ❖ Write a fingerplay
- ❖ Timeline calendar
- ❖ Introduction centers
- ❖ Mystery activities
- ❖ Share the news
- ❖ I have who has name game
- ❖ Community-building activities
- ❖ Tally
- ❖ Weather graphing

The table below provides descriptive data for the teacher-reported implementation frequency for all activities in the dataset. Overall, make-believe play, attention-focusing, and introduction activities exhibited the highest means for teacher-reported implementation. This means that teachers frequently implemented the make-believe play activity, which aligns with the Tools developers' expectations in the manual (Leong & Bodrova, 2011). For descriptive implementation frequency data for each activity block, please see section 12.3.

Table: Descriptive statistics for teacher-reported Tools activity implementation

Activity	Min	Max	Mean	SD
Attention-focusing	1.5	3	2.90	.28
Attribute game	0	3	1.73	.75
Buddy reading	0	3	2.13	.81
Community building	1.5	3	2.25	.40
Complete and continue	0	2.5	.96	.76
Elkonin boxes	0	3	1.59	.78
Freeze game	1	3	2.66	.46
Freeze on number	0	3	1.55	.90
Graphics practice	0	3	2.12	.50
I have colors	.5	3	1.81	.49
I have letters	.5	3	1.87	.56
I have names	0	3	1.49	1.00
I have numbers	.5	2.5	1.73	.52
I have shapes	1	2.5	1.75	.46
Introduction to centers	1	3	2.20	.59
Making collections	.5	2.5	1.96	.45
Make a rhyme	0	3	1.73	.64
Math memory	.5	2.5	1.74	.50
Make-believe play centers	1	3	2.80	.42
Make-believe play cleanup	0	3	2.69	.71
Make-believe play planning	0	3	2.67	.66
Make-believe play practice	1	3	2.30	.67
Message of the day	1	3	2.90	.36
Mouse trap	0	3	.95	.87
Mr. Wolf	0	3	.97	.85
Mystery activities	2	3	2.93	.23
Number follow the leader	0	2.5	1.11	.66
Numeral games	.5	3	1.83	.69
Numberline hopscotch	0	3	1.70	.75
Partner freeze	0	3	1.85	.81
Patterns and manipulatives	.5	3	1.97	.66
Pattern movement	.5	3	1.76	.63

Pretend transitions	.5	3	2.50	.58
Puzzles with manipulatives	.5	3	2.46	.67
Remember and replicate	0	2.5	1.61	.54
Science eyes	0	2.5	1.75	.43
Share the news	1.5	3	2.74	.43
Story lab activities	2	3	2.77	.36
Takeaway sounds	0	3	1.59	.86
Tally	.5	3	1.75	.66
Timeline calendar	1.5	3	2.95	.24
Two step freeze	0	3	1.45	.81
Venger drawings	.5	3	1.91	.67
Weather graph	2.5	3	2.96	.14
Write a fingerplay	0	3	1.43	.74

The table below provides descriptive data for the standardized fidelity of implementation for all activities in the dataset. Because the data has been standardized, all variables have a mean of zero and standard deviation of 1. However, the minimum and maximum values provide a sense of the range. That is, one teacher implemented community-building activities with a very high level of fidelity (z-score = 5.07); by contrast, one teacher implemented make-believe play planning with very low levels of fidelity (z-score = -3.53).

Table: Descriptive statistics for standardized fidelity of activity implementation

Activity	Min	Max	Mean	SD
Activities throughout the day	-2.50	2.06	0	1
Attention-focusing activities	-1.48	2.35	0	1
Attribute game	-2.28	2.49	0	1
Class schedules	-.54	2.82	0	1
Community building activities	-1.98	5.07	0	1
Complete and continue	-1.52	4.87	0	1
Elkonin boxes	-2.26	5.10	0	1
Freeze game	-2.19	1.47	0	1
Freeze on number	-1.35	2.87	0	1
Literacy activities	-2.81	1.86	0	1
I have who has colors	-1.89	3.58	0	1
I have who has letters	-1.33	2.80	0	1
I have who has names	-.35	3.72	0	1
I have who has numbers	-2.61	2.57	0	1
I have who has shapes	-.71	2.67	0	1
Buddy reading	-2.86	1.69	0	1

Make-believe play centers	-2.50	1.70	0	1
Make-believe play clean up	-1.93	1.34	0	1
Make-believe play planning	-3.53	1.09	0	1
Make-believe play practice	-.87	3.34	0	1
Making collections	-.67	2.49	0	1
Math memory	-.58	2.53	0	1
Science eyes	-1.74	2.90	0	1
Message of the day	-3.43	1.31	0	1
Mystery activities	-2.87	1.68	0	1
Numberline hopscotch	-2.82	2.87	0	1
Numerals game	-2.16	2.86	0	1
Partner freeze	-2.15	1.85	0	1
Pattern movement game	-.58	3.06	0	1
Patterns with manipulatives	-1.93	4.14	0	1
Pretend transition	-.92	2.60	0	1
Pattern movement	-3.33	3.68	0	1
Remember and replicate	-.72	2.08	0	1
Science eyes	-.60	3.34	0	1
Share the news	-2.50	1.51	0	1
Story lab connections	-.76	3.00	0	1
Story lab stories	-1.35	3.21	0	1
Story lab small group	-1.71	3.69	0	1
Story lab learning facts	-.41	3.66	0	1
Story lab predictions	-1.08	5.25	0	1
Story lab visual literacy	-.47	3.32	0	1
Story lab vocabulary	-.59	3.25	0	1
Take away sounds	-1.00	4.34	0	1
Tally	-1.72	3.19	0	1
Timeline calendar	-1.77	1.56	0	1
Two-step freeze	-2.58	2.43	0	1
Venger drawing	-2.17	2.17	0	1
Weather graphing	-2.77	.75	0	1

Appendix K: Analysis of all Tools activities

This appendix contains the results of the Tools activities not presented in section 12.3. That section shared only the results from the attention-focusing, math, and make-believe play blocks, whereas this appendix includes all the remaining activities from the literacy, introduction, and science blocks (see the table below).

Table: Intercept and slope estimates for literacy, introduction, and science activities across the three outcome measures

	Executive function		SAR		CFBRS	
	I	S	I	S	I	S
Buddy reading	.02	-.05	-.02	-.05	-.03	.05
Community building	-.19***	.11	-.06	.10	-.16**	.18*
Elkonin boxes	.00	-.12*	-.03	-.17*	-.06	-.02
Graphics practice	-.06	-.03	-.10*	.07	-.10	.06
I have who has letters	-.13*	.01	-.03	-.08	-.11 ⁺	.06
I have who has names	-.07	-.05	-.02	-.09	-.14 ⁺	.09
Introduction to centers	.01	-.04	-.06	-.05	-.09	.17 ⁺
Make a rhyme	-.08	-.03	-.05	-.07	-.03	.01
Message of the day	-.04	-.05	-.08 ⁺	.05	.00	-.17 ⁺
Mystery activities	-.02	.08	.05	-.05	.14*	-.12*
Number follow the leader	.01	-.08	-.03	-.09	.04	-.11
Science eyes	-.06	-.02	-.08	.04	-.06	-.02
Share the news	-.03	-.10*	-.12**	.06	-.00	-.16
Story lab activities	-.06	-.02	-.17***	.16*	-.03	-.10
Takeaway sounds	-.01	-.09	-.04	-.11	-.06	.07
Tally	-.03	-.09	-.06	-.12	.01	-.09
Timeline calendar	.01	-.03	-.03	.03	.00	-.06
Weather graph	.08*	.00	.13**	-.08	.07 ⁺	.09
Write a fingerplay	.01	.02	-.02	-.07	-.00	.04

As observed in the table above, only the weather graphing activity, where teachers and students create a weekly bar graph to capture the daily weather, predicts significantly higher self-regulation skills. That association is consistent across the intercepts of all three outcome measures, and it is not counterbalanced by any negative associations with the slope. Besides

weather graphing, though, all other activities follow the pattern of null or negative associations with children's self-regulation skills.

Whereas the table above shows the findings from the teacher-reported implementation frequency measure, the table below shows the sensitivity results for the full set of researcher-reported Tools implementation fidelity variables. Specifically, the table below indicates the estimates for each activity's beta coefficient on the intercept and slope of all three self-regulation outcome measures.

As with the main activity analyses, most of the fidelity variables yielded null or negative associations with each self-regulation outcome. However, one activity, Story Lab, exhibited a significantly positive beta coefficient ($b = .17, p < .05$) for the researcher-reported self-regulation growth through first grade.

In addition to Story Lab's significantly positive association, five of the fidelity-based activities yielded marginal associations ($.05 < p < .10$) that were positive without any counterbalancing negative sign: Attribute game, Elkonin boxes, I have who has colors, Making collections, and Take away sounds. Beyond those, none of the other activities yielded an exclusively positive association with self-regulation; that is, even where a positive association was observed with the intercept, the slope had a negative sign to counterbalance the initial positive association, and vice versa.

Table: Intercept and slope estimates for all researcher-observed implementation fidelity activity variables across the three outcome measures

	EF		CFRBS		SAR	
	Intercept	Slope	Intercept	Slope	Intercept	Slope
Activities throughout the day	.04	-.13*	.04	-.12	-.11	-.01
Attention-focusing activities	.13*	-.13 ⁺	.09	-.06	-.09	-.10
Attribute game	.07	-.02	.10 ⁺	-.08	-.02	.06
Class schedules	.02	-.19**	-.12 ⁺	.08	-.08	-.08
Community building activities	.02	-.04	.02	.02	-.01	.02
Complete and continue	-.07	.04	.02	-.08	-.03	.13
Elkonin boxes	-.04	.11*	-.01	.14 ⁺	.05 ⁺	-.05
Freeze game	.01	.00	-.04	-.01	.01	-.03
Freeze on number	.00	-.11 ⁺	-.02	-.15 ⁺	-.06	.13
Literacy activities	.03	-.16 ⁺	-.03	-.01	-.11 ⁺	-.05
I have who has colors	.08 ⁺	-.01	.05	.02	-.05	.14 ⁺

I have who has letters	-.09	.00	-.17*	.11	-.01	-.03
I have who has names	.05	-.11	-.03	.01	-.02	.00
I have who has numbers	.01	.00	.01	.15	-.09*	-.01
I have who has shapes	-.03	.03	-.15*	.12	-.01	.15
Buddy reading	.04	-.17 ⁺	-.06	.02	-.10	-.03
Make-believe play centers	.00	-.07	-.07	.02	-.13*	.20*
Make-believe play clean up	.10	-.18**	-.02	-.07	-.09	.03
Make-believe play planning	.06	-.03	.06	-.05	-.08 ⁺	.12 ⁺
Make-believe play practice	.02	-.10	.06	.00	-.16*	.15
Making collections	-.01	.11	-.01	.04	.10 ⁺	-.01
Math memory	.12 ⁺	-.13 ⁺	.04	.05	.09 ⁺	-.19**
Science eyes	.09	-.07	.02	.01	-.08	.12
Message of the day	-.08 ⁺	-.05	-.12	.03	-.13*	.17*
Mystery activities	.04	-.07	.04	-.09	-.03	-.03
Numberline hopscotch	.14**	-.17*	.04	.05	-.02	.14
Numerals game	-.01	-.01	.04	-.07	-.06	.02
Partner freeze	.01	.01	.01	.01	-.04	-.11
Pattern movement game	.03	.00	.02	-.11	-.08 ⁺	-.04
Patterns with manipulatives	-.02	.01	.03	.01	.01	-.02
Pretend transition	-.03	.03	.11	-.12	-.10 ⁺	.20*
Remember and replicate	.04	-.05	-.01	.01	-.04	-.13
Share the news	.04	-.10 ⁺	.01	-.10	-.14**	.08
Story lab	.04	-.06	.03	.17	-.08	.17*
Take away sounds	-.05 ⁺	.00	-.07	-.05	.01	.16 ⁺
Tally	.01	-.05	.03	.03	-.08	.03
Timeline calendar	-.08	-.05	-.03	-.03	-.10	.14
Two-step freeze	-.03	-.01	-.12 ⁺	.16*	.00	-.16*
Venger drawing	-.08	.09	-.08	.10	-.08	.11
Weather graphing	.06	-.07	-.02	.09	.01	-.04

To summarize, the majority of the researcher-observed fidelity findings (107 of the 120 total beta coefficients, or 89.2%) exhibited null or negative associations between Tools activities and children's self-regulation. This outcome generally replicates the findings observed for the teacher-reported implementation frequency data observed in Chapter Twelve as well as in the table above.

Appendix L: Multilevel models for change

Multilevel models for change, also called latent growth models, are underpinned by two levels of equations. These equations involve the following four variable types: a time measure, time-varying covariates, time-invariant covariates, and an outcome measure (Cillessen & Borch, 2006). In this study, the time measure consists of the four self-regulation data collection sweeps (beginning and end of pre-kindergarten, end of kindergarten, and end of first grade). Time-invariant variables do not change across time (e.g., child gender). Time-varying variables (e.g., child's age) do change across time. The outcome measure is children's self-regulation skills.

The level-1 equation includes variables that change over time, which, in this analysis, is only child age. The level-2 equation includes variables that do not change over time, such as children's gender. When combined, the level-1 and level-2 models illustrate the within-person and between-person variability in self-regulation growth over time. The level-1 and level-2 equations are explained below.

MLM for change: Level-1

The level-1 model estimates the within-person change over time on the outcome variable as well as the effect of time-varying predictors on this development (Cillessen & Borch, 2006). Using Singer and Willett's (2003) guidance for notation, the level-1 equation is as follows:

$$Y_{ij} = \pi_{0i} + \pi_{1i}TIME_{ij} + \epsilon_{ij}$$

In the above equation, Y_{ij} is the predicted score on self-regulation for a given child i at time point j . The initial status, π_{0i} , is the value of the self-regulation for child i when time is zero (i.e., the first data collection point at the beginning of pre-kindergarten). The growth rate, π_{1i} , indicates the rate of change in self-regulation for child i . Finally, the error term, ϵ_{ij} , is the within-person error for child i at time j (Singer & Willett, 2003). Because this model contains only the initial status and growth rate of self-regulation without including any time-varying predictors or covariates, this model is called the unconditional growth model.

Testing the unconditional growth model is an important prerequisite for further analysis because it confirms whether there is sufficient within-person variability to warrant multilevel modeling (Singer & Willett, 2003). If people do exhibit significantly different initial statuses and growth rates, then other time-varying predictors can be added to the level-1 model. When the level-1 model incorporates time-varying predictors, it is adapted to resemble the equation below:

$$Y_{ij} = \pi_{0i} + \pi_{1i}TIME_{ij} + \pi_{2i}X_{2ij} + \pi_{3i}X_{3ij} + \varepsilon_{ij}$$

In the equation above, the variables labeled “X” represent the effect of a time-varying predictor (e.g., child age) on self-regulation. All the other terms in the equation have the same meaning as those described above for the unconditional growth model. The level-1 model investigates only within-person change in self-regulation over time.

MLM for change: Level-2

The level-2 model explores between-person variability. The predictors of the level-1 model become the outcome variables at level-2. Specifically, the initial status (π_{0i}), growth rate (π_{1i}), and time-varying predictors (π_{2i} , π_{3i} , π_{ni}) are now outcome variables in the series of level-2 equations, which have their own intercept, error, and time-invariant predictors as illustrated in this series of equations:

$$\pi_{0i} = \gamma_{00} + \gamma_{01}M_i + \zeta_{0i}$$

$$\pi_{1i} = \gamma_{10} + \gamma_{11}M_i + \zeta_{1i}$$

$$\pi_{2i} = \gamma_{20} + \gamma_{21}M_i + \zeta_{2i}$$

$$\pi_{ni} = \gamma_{n0} + \gamma_{n1}M_i + \zeta_{ni}$$

In the level-2 equations, the time-invariant variable of gender is entered as a predictor. In the equations above, the term “M” represents male, which was coded as “1” for this analysis (females = 0). The level-2 intercepts (γ_{00} , γ_{10} , γ_{20} , and γ_{n0}) represent the values of the level-1 parameters when the time-invariant predictor (i.e., gender) is equal to zero (i.e., when the

child is male). Given that females are coded as zero and males as one in the PRI dataset, the models indicate how much the level-1 parameters change for males relative to females. The ζ_{ni} error terms are indicative of individual variability in the level-1 parameters that are not explained by the level-2 predictors (Cillessen & Borch, 2006).

Appendix M: Sample Mplus model specification syntax

This appendix contains sample Mplus syntax to show the latent growth model specifications used in study two (Section IV of the dissertation). Each section below contains a bolded annotation below the line of code. I have added those annotations here to explain the function of the line of code in Mplus.

TITLE: SAR ON MBP ;

The “TITLE” command gives a name to the Mplus syntax file.

DATA: FILE IS

/Users/abaro2/Documents/DPhil Imputation/Imputed datasets/28Jan_TeaNarlist.dat ;

The “DATA” command indicates the location of the data file on the computer, which must be in a .dat or .txt format for Mplus to read.

TYPE = IMPUTATION ;

The “TYPE = IMPUTATION” command signifies to Mplus that the software must run the analyses with the five imputed datasets. See Chapter Eleven for an explanation of multiple imputation as the missing data treatment for this thesis.

VARIABLE: NAMES ARE

Attfocus AttGame BudRead ComBuild CompCont ElkBox FreeGame FreezNum

GraPrac HaveCol HaveLet HaveName HaveNum HaveShap IntroCent MakeColl

MakeRhym MathMem MBPCent MBPClean MBPPlan MBPPrac MessDay MousTrap

MrWolf MystAct NumFolLe NumGames NumHop ParFreez PatManip PattMov

PreTran PuzManip RemRepl SciEyes ShareNew SLAct TakSound Tally

TimeCal TSFreeze Venger WeathGra WriteFpl PosBehav BehavRem ChoralRes

PairedAc Scaffold PrivSpee AgeT1 AgeT2 AgeT3 AgeT4 T1T2_int T2T3_int T3T4_int

CB_t1 CB_t2 CB_t3 CB_t4 DCCS_t1 DCCS_t2 DCCS_t3 DCCS_t4

CD_t1 CD_t2 CD_t3 CD_t4 HTKS_t1 HTKS_t2 HTKS_t3 HTKS_t4

```
PT_t1 PT_t2 PT_t3 PT_t4 S1_t1 S2_t1 S3_t1 S4_t1 S5_t1
S1_t2 S2_t2 S3_t2 S4_t2 S5_t2 S1_t3 S2_t3 S3_t3 S4_t3 S5_t3
S1_t4 S2_t4 S3_t4 S4_t4 S5_t4 C06_t1 C27_t1 C29_t1 C31_t1 C32_t1
C06_t2 C27_t2 C29_t2 C31_t2 C32_t2 C06_t3 C27_t3 C29_t3 C31_t3 C32_t3
C06_t4 C27_t4 C29_t4 C31_t4 C32_t4 ID Cohort SchID ELL Gender IEP TeachID ;
```

The **“VARIABLES”** section assigns the correct variable names to each column in the .dat data file, which does not have any data names written into it.

```
USEVAR =
S1_t1 S2_t1 S3_t1 S4_t1 S5_t1 S1_t2 S2_t2 S3_t2 S4_t2 S5_t2
S1_t3 S2_t3 S3_t3 S4_t3 S5_t3 S1_t4 S2_t4 S3_t4 S4_t4 S5_t4
ELL Gender IEP Age_c MBP ;
```

The **“USEVAR”** section informs Mplus which variables will be used in this analysis. Any variables in the subsequent sections that are not mentioned here will result in an error message.

```
IDVARIABLE = ID ;
CLUSTER = TEACHID ;
```

The **“IDVARIABLE”** command informs Mplus that the identification variable at level one, or the child-level, is each child’s ID number, which was assigned by the PRI researchers. The **“CLUSTER”** command informs Mplus that the second-level in the multilevel analysis is the teacher, or classroom level. This adjusts the standard errors to account for data dependency among children who have the same teacher.

```
DEFINE:
Age_c = AgeT1 - 54 ;
MBP = MEAN (MBPCent MBPPlan MBPPrac) ;
```

The **“DEFINE”** command can be used for variable creation in Mplus. The **“Age_c”** variable is the centered age variable described in Appendix N. The **“MBP”** variable is the composite make-believe activity block variable that will be used to predict children’s self-regulation skills in the analysis below.

```
ANALYSIS: ESTIMATOR = MLR ;
```

The maximum-likelihood estimator (MLR) is specified above to use an iterative optimization algorithm that seeks parameter estimates that are ‘maximally likely’ to have produced the observed data matrix.

TYPE = COMPLEX ;

The ‘Type = Complex’ command clusters standard errors at the classroom level, which was designated as the cluster in the syntax noted above. Again, this process reduces Type I error rates by correctly estimating standard errors through accounting for data dependency among children in the same classroom.

MODEL:

! Variable covariances to estimate patterns for configural invariance;

S1_t1 WITH S1_t2 S1_t3 S1_t4 ;

S1_t2 WITH S1_t3 S1_t4 ;

S1_t3 WITH S1_t4 ;

S2_t1 WITH S2_t2 S2_t3 S2_t4 ;

S2_t2 WITH S2_t3 S2_t4 ;

S2_t3 WITH S2_t4 ;

S3_t1 WITH S3_t2 S3_t3 S3_t4 ;

S3_t2 WITH S3_t3 S3_t4 ;

S3_t3 WITH S3_t4 ;

S4_t1 WITH S4_t2 S3_t3 S3_t4 ;

S4_t2 WITH S4_t3 S3_t4 ;

S4_t3 WITH S4_t4 ;

S5_t1 WITH S5_t2 S5_t3 S5_t4 ;

S5_t2 WITH S5_t3 S5_t4 ;

S5_t3 WITH S5_t4 ;

This section checks inter-item covariances to assess whether the five SAR items have similar correlations with one another across time points. This is the first part of

measurement invariance testing for configural invariance as described in Appendix I. Namely, the 'WITH' command causes Mplus to return inter-item correlations. By checking the consistency of correlations among items across time points, we can investigate configural invariance.

!Fixed factor loadings ;

SR_1 BY S1_t1@1

S2_t1 S3_t1 S4_t1 S5_t1 (2-5) ;

SR_2 BY S1_t2@1

S2_t2 S3_t2 S4_t2 S5_t2 (2-5) ;

SR_3 BY S1_t3@1

S2_t3 S3_t3 S4_t3 S5_t3 (2-5) ;

SR_4 BY S1_t4@1

S2_t4 S3_t4 S4_t4 S5_t4 (2-5) ;

The second part of the model defines the latent self-regulation construct in terms of the observed self-regulation assessor rating (SAR) scores. In the model specification, the factor loadings for each of the five SAR items are constrained to be equal across all four time points. This added constraint defines the weak invariance model described in Appendix I.

! Fixed intercepts ;

[S1_t1 S1_t2 S1_t3 S1_t4] (6) ;

[S2_t1 S2_t2 S2_t3 S2_t4] (7) ;

[S3_t1 S3_t2 S3_t3 S3_t4] (8) ;

[S4_t1 S4_t2 S4_t3 S4_t4] (9) ;

[S5_t1 S5_t2 S5_t3 S5_t4] (10) ;

The third part of the model is the final element of measurement invariance testing. Namely, each of the latent item means are constrained to be equal across time, which is the requirement for strong invariance models (see Appendix I). Thus, the first three parts of the model involve the confirmatory factor analysis and measurement invariance models. Given the excellent fit ($CFI > .95$ and $RMSEA <$

.08), we can assume measurement invariance holds and proceed to the latent growth model specification below.

```
! Latent growth model ;
```

```
i s | SR_1@-1 SR_2@0 SR_3@2 SR_4@4 ;
```

The syntax above defines the latent intercept and slope parameters in terms of the latent self-regulation constructs. The ‘@’ signs in the model specification enable slope parameterization as described in the final section of Chapter Eleven. One note on all growth models in this study is that the latent intercept and slope parameters are themselves defined by latent self-regulation constructs. In traditional growth modeling, latent intercept and slopes are often defined by observed manifest indicators (Singer & Willett, 2003).

By contrast, the intercepts and slopes here are defined by latent self-regulation constructs, which are defined by manifest indicators (i.e., task-based executive function measures and informant-report self-regulation measures). Thus, the model specification here is more complex than a traditional growth model, but it still fits within the latent growth model framework explicated in Wu et al. (2010) and is recommended by Little (2013).

By defining the latent intercept and slope in terms of latent self-regulation constructs, we are employing the “multiple indicator multilevel growth model” framework outlined in Wu et al. (2010). That paper explains the benefits of defining latent intercepts and slopes in terms of latent constructs, which in turn can derive from either standardized or standardized manifest indicators.

The central benefit of this approach is that the latent self-regulation constructs partition variance into true variance and error variance. By accounting for the error variance, or measurement error, the resultant latent estimates more accurately capture the underlying construct (i.e., self-regulation).

Consequently, the latent growth model’s intercept and slope estimates are more accurate because they are based on latent constructs that have already accounted for measurement error. Despite the intuitive appeal of this latent growth model approach, it has yet to become widespread practice in the social sciences (Wu, Liu, Gadermann, & Zumbo, 2010).

```
i s ON ELL Gender IEP Age_c MBP ;
```

The final line of Mplus code produces the conditional growth model. That is, the latent self-regulation intercept and slope are conditioned, or statistically predicted, by the vector of covariates (i.e., child gender, special education status, English language learner status, and centered age) as well as the Tools make-believe play composite variable.

Appendix N: Various treatments of time variables

Children's neuropsychological maturation emerges as a consistent predictor of children's self-regulation development. Thus, controlling for the effect of time on children's self-regulation skills was a critical issue in these analyses. In order to most effectively control for time, four approaches were considered. Ultimately, the most parsimonious approach (i.e., controlling for children's centered age in months at the beginning of pre-kindergarten) emerged as the most sensible solution. That said, the three other attempted approaches are explained in the sections below.

Second approach: Age as a time-varying covariate

Instead of simply controlling for children's age at time one (i.e., the beginning of pre-kindergarten), I also attempted controlling for children's age at each of the four time points. In so doing, time was entered into the latent growth model as a time-varying covariate; that is, at each time point, children's age value was different, which explains its designation as a 'time-varying' covariate. The time-varying Mplus syntax is the following:

SR_1 ON AgeT1 ;

SR_2 ON AgeT2 ;

SR_3 ON AgeT3 ;

SR_4 ON AgeT4 ;

Although age at each time point significantly predicts ($p < .001$) self-regulation skills, the age variables are highly collinear. In fact, each of the age variables is correlated above $r = .95$, which all exceed Field's (2013) guideline of .9 for problematic multi-collinearity. Thus, only the first age variable was kept in the model, and it was used to predict children's self-regulation intercept and growth rate. The resultant Mplus syntax is below:

I S ON ELL Gender IEP AgeT1 ;

SR_2 ON AgeT2 ;

SR_3 ON AgeT3 ;

SR_4 ON AgeT4 ;

Once time one age was entered as a predictor of the intercept and slope, the other age variables all became insignificant predictors of self-regulation. This is because the age at time one explains significant variance in self-regulation growth over time, whereas the subsequent age variables are all highly correlated with age at time one; thus, they do not explain additional variance above and beyond time one age. Given the multi-collinearity issues and the insignificant predictive value from adding more age variables, only centered age at time one was retained in the model.

Third approach: Controlling for age and test intervals

Although controlling for age as a time-varying covariate was imprudent due to collinearity, I tried controlling for intervals between testing administrations as a time-varying covariate. That is, at each time point, the temporal interval between the first self-regulation testing and the second testing period (i.e., from fall to spring of pre-kindergarten) was inserted as a control variable. The sample Mplus syntax is as follows:

I S ON ELL Gender IEP AgeT1 ;

SR_2 ON T1T2_Int ;

SR_3 ON T2T3_Int ;

SR_4 ON T3T4_Int ;

This approach could be considered sensible because children were tested at slightly different times over the four-year period. Once again, these data were collected over four time points: 1) fall of pre-kindergarten, 2) spring of pre-kindergarten, 3) spring of kindergarten, and 4) spring of first grade. However, given the scope of the project, the testing of individual children took place across the span of several weeks per time point. Thus, some children

were tested weeks before others, which could affect their self-regulation scores. Moreover, the intervals for each child were not consistent across time points; that is, a child tested relatively early in the time point one distribution may have been tested relatively late in the time point two distribution.

Without accounting for that variability, it might be concluded that one child's self-regulation capacity grew more substantially than another child's; however, the true explanation may be that one child simply had more time between testing administrations, so his or her growth score appears more noteworthy than it actually is. Because self-regulation capacity develops so rapidly during early childhood (Bronson, 2000; Kopp, 1982; Moilanen, Shaw, Dishion, Gardner, & Wilson, 2009), temporal analysis of its growth must be very precise.

Although this approach may seem to make sense, the test interval variables did not predict additional variance beyond that already explained by age. That is, all of the test intervals predictors were non-significant. Thus, I ultimately omitted them from the model to achieve parsimony given that the centered age variable from time point one predicted self-regulation development better than the other approaches.

Fourth approach: Random effects for time

The most sophisticated approach I attempted was to include random effects for time. This approach is called "individually-varying times" in the Mplus package (Muthén & Muthén, 2012, p. 130). The approach relies upon stochastic differential equations, which integrate each child's exact age at data collection into the model to compute maximally accurate parameter estimates for each child (Voelkle, Oud, Davidov, & Schmidt, 2012). Thus, similar to the third approach described above, this approach tries to correct for subtle differences in data collection intervals across children.

While the third approach described above uses test intervals, this approach uses the child's exact age in months to compute the parameter estimates. As with the test intervals approach above, the random effects for time model did not fit any better than the most parsimonious model with the fixed effect for centered age. In fact, the model fit slightly less well. Thus, given considerations of parsimony, this approach was also abandoned.