Addressing an Old Issue from a New Methodological Perspective: A Proposition on How to Deal with Bias due to Multilevel Measurement Error in the Estimation of the Effects of School Composition

Ioulia Televantou

Jesus College, University of Oxford

Thesis submitted to the University of Oxford for the degree of

DPhil in Education

Trinity Term, 2014

Dedication

I dedicate this thesis to my grandfather, Michael.

Acknowledgements

I would first like to thank my supervisor Herb Marsh for all his interest in my work and for the guidance that he has been constantly providing to me so far. I deeply appreciate the fact that although he has been miles away from me most of the time over the last couple of years, he being in Australia and I being in Cyprus, he still continued to monitor my work with enthusiasm.

I would also like to take this opportunity to express my gratitude to my colleagues Benjamin Nagengast and John Fletcher for the time they devoted to discussing my work and theirs. I am equally grateful to Lars Malmberg, not only for the feedback that he gave me at the transfer of status stage, together with Gordon Stanley, but also for co-supervising me after Herb's return to Australia. Furthermore, I would like to thank my college advisor, Kathy Sylva, for all her advice and support during my years in Oxford.

Moreover, I owe a debt of gratitude to Chrystalla Demetriades; without whom I would not be academically at the stage I am now; to Leonidas Kyriakides and to Peter Tymss and the CEM center in Durham; the data that they have provided for my research allowed me to explore my research questions in many different ways. I am also grateful to Kit Tai Hau and to David Andrich and Joshua McGrane for their assistance with my data analyses during their time as visiting scholars at Oxford; to Pam Sammons and Jo-Anne Baird for the constructive feedback that they gave on my Confirmation of Status Report.

It would be a great omission not to thank some special friends – people who were there to help me whenever I needed them providing me with practical, emotional and intellectual support: Evdokia Pittas, Karolina Retalis, Charalambos Themistocleous, Peter Edelsbrunner, Ioannis Kyriakides, Maria Evangellou, Zorka Sljivancanin, Savina Joseph, Eleni and Theodora Lymboura and many others.

Last but not least, I owe special thanks to my much loved parents, Anastasios and Antonia and to my brother Michael for their love, patience and support.

Abstract

With educational effectiveness studies, school-level aggregates of students' characteristics (e.g. achievement) are often used to assess the impact of school composition on students' outcomes – school compositional effects. Empirical findings on the magnitude and direction of school compositional effects have not been consistent. Relevant methodological studies raise the issue of under-specification at level 1 in compositional models - evident when the student-level indicator on which the aggregation is based is mis-measured. This phenomenon has been shown to bias compositional effect estimates, leading to misleading effects of the aggregated variables – phantom compositional effects. My thesis, consisted of three separate studies, presents an advanced methodological framework that can be used to investigate the effect of school composition net of measurement error bias.

In Study 1, I quantify the impact of failing to account for measurement error on school compositional effects as used in value added models of educational effectiveness to explain relative school effects. Building on previous studies, multilevel structural equation models are incorporated to control for measurement error and/or sampling error. Study 1a, a large sample of English primary students in years one and four (9,059 students from 593 schools) reveals a small, significant and negative compositional effect on students' subsequent mathematics achievement that becomes more negative after controlling for measurement error. Study 1b, a large study of Cyprus primary students in year four (1694 students in 59 schools) shows a small, positive but statistically significant effect that becomes non-significant after controlling for measurement error.

Further analyses with the English data (Study 2), demonstrates a negative compositional effect of school average mathematics achievement on subsequent mathematics self-concept – a Big Fish Little Pond Effect (BFLPE). Adjustments for measurement and sampling error result in more negative BFLPEs. The originality of Study 2 lies in verifying BFLPEs for students as young as five to eight/nine years old. Bridging the findings related to students' mathematics self-concept (Study 2) and the findings on students' mathematics achievements (Study 1a), I demonstrate that the prevalence of BFLPEs with the English data partly explains the negative compositional effect of school average mathematics achievement on students' subsequent mathematics achievement.

Lastly, in Study 3 I consider an alternative approach to school accountability to conventional value added models, namely the Regression Discontinuity approach. Specifically, I use the English TIMSS 1995 primary (years four and five) and secondary (years eight and nine) data to investigate the effect of one extra year of schooling on students' mathematics achievement and the variability across schools in their absolute effects. The extent to which school composition, as given by school average achievement, correlates with schools' added-year effects is addressed. Importantly the robustness of the RD estimates to measurement error bias is demonstrated.

My findings have important methodological, substantive and theoretical implications for on-going debates on the school compositional effects on students' outcomes, because nearly all previous research has been based on traditional approaches to multilevel models, which are positively biased due to the failure to control for measurement error.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ACH	Achievement
ASC	Academic Self Concept
BFLPE	Big-Fish-Little-Pond-Effect
CEM	Centre of Evaluation and Monitoring
CFA	Confirmatory Factor Analysis
CFI	Confirmatory Fit Index
CVA	Contextual Value Added model
CTT	Classical Test Theory
E-CVA	English Contextual Value Added Model
EER	Educational Effectiveness Research
ESF	"Dynamics of Educational Effectiveness" project
GPA	Grade Point Average
ICC	Intraclass Correlation Coefficient

L1	Individual-level, level 1
L2	Group-level, level 2
MACH1	Mathematics achievement in year one
MACH4	Mathematics achievement in year four
MCAR	Missing Completely at Random
MI	Multiple Imputation
MLM	Multilevel Modelling
ML-SEM	Multilevel Structural Equation Models
MNAR	Missing Not at Random
MSC1	Mathematics self-concept in year one
MSC4	Mathematics self-concept in year four
PIPS	Performance Indicators at Primary School project
RD	Regression Discontinuity
RDD	Regression Discontinuity Design
SER	School Effectiveness Research
TIMSS-95	Third International Mathematics and Science Study

Chapter 1: Introduction

The development of educational research requires sophisticated methodologies: complex methodological tools allow classic substantive topics to be explored in more detail and provide insights into issues not approachable in the past; methodological-substantive synergies (Marsh and Hau, 2007). My thesis uses cutting-edge methodological developments from the field of educational psychology and self-concept research (Lüdtke, Marsh, Robitzsch, Trautwein, Asparouhov, and Muthén, 2008; Marsh, Lüdtke, Robitzsch, Trautwein, Asparouhov, Muthén and Nagengast, 2009) and applies them in educational effectiveness to address a highly debated topic within the field–namely the potential effect of school composition on students' educational outcomes.

Comprised of three separate studies, the overarching aim of the present investigation is to introduce a new methodological perspective in the assessment of the school composition effect in educational effectiveness research – one that can deal with bias due to multilevel measurement error in relevant statistical estimates. All three studies build on and expand previous work on the effect of school composition by exploring the role of measurement error in its estimation using different data sets.

In what follows, I develop the rationale underlying my research. I begin with an overview of the methodological and substantive issues addressed in my thesis, defining terms and indicating where I build on existing literature. I then summarise the main objectives of the implemented work, stemming from methodological weaknesses identified in relevant studies. Finally, I outline the content chapters in order to provide the overall organization of the thesis.

1.1 Compositional Effects: Definitions and Terminology

In educational settings it is often of interest to investigate whether school, classroom or teacher (level 2, higher-level, group-level) characteristics contribute to the prediction of students' (level 1, lower-level, individual-level) characteristics, for example, their achievement or their self-concept, over and above what can be explained by pre-existing differences. In many studies, the higher-level constructs are formed by aggregating the lower-level characteristic (e.g. the students' achievement or socio-economic status) at the group-level (e.g. the school, the teacher or the classroom). The question is then whether the obtained higher-level characteristic has an effect on the individual-level variable on which aggregation is based (Lüdtke et al., 2008). Such effects are commonly referred in the literature as "compositional" effects (e.g. Nash, 2003; Harker and Tymms, 2004).

The term "contextual" has also been widely used in the literature (Boyd and Iverson, 1979; Bryk and Raudenbush, 2002) to denote the effect of an aggregate variable over and above that of the corresponding individual-level characteristic. This terminology has been mainly adopted by research within the field of educational psychology. Although some of these studies are particularly relevant to the present investigation (e.g. Lüdtke et al.,2008; Marsh, Lüdtke et al., 2009; Lüdtke, Marsh et al., 2011), I have chosen throughout my thesis to use the term "compositional", in accordance with recommendations made by Harker and Tymms (2004).

It is claimed that the word "context" or "contextual", particularly in the educational effectiveness paradigm, encompasses a wider meaning than the term "compositional. It is often used to denote the wider properties of the pupil background and other environmental factors that may affect students' outcomes (e.g. the neighbourhood context or other characteristics of the educational community in which the student is taught). In this respect, the effect of the school (or classroom) composition is just one component of this contextual effect.

1.2 Multilevel Modelling: The Conventional Approach to Compositional Effects Analysis

Compositional effects are typically detected by statistical analysis (see also Nash, 2003), namely multilevel modelling. Multilevel modelling effectively takes into account the hierarchical structure in educational data (e.g. students nested within schools), adjusting for the inter-independence of the observations within the same primary sampling unit (for example, the school). It allows the assessment of relationships between variables measured at two different levels simultaneously (e.g. the effects of school-level variables on individual-level outcomes). Ignoring the group-level and conducting the analysis at the individual-level could lead to underestimation of the standard errors and result in invalid statistical test results especially for compositional variables. The reverse solution - aggregating the data to the higher-level and ignoring the individual-level – may lead to ecological fallacies (Robinson, 1950).

The special class of multilevel models that are used to assess the magnitude and the direction of compositional effects is often referred to in the literature as compositional models (see Harker and Tymms, 2004; and also the section in the literature review on terminology). With compositional models the criterion is typically individual-level outcomes of interest (e.g. students' academic achievement, students' self-concept); these are regressed on the individual-level predictor (e.g. students' prior achievement) and on the corresponding higher-level aggregate (e.g. school-level prior achievement). Hence the effect of the aggregated variable on the individual outcome is the compositional effect. The effect of the individual-level variable on which aggregation is based can also be of interest in compositional effects models and is referred to as the within-group effect. Despite the advantages associated with the use of conventional multilevel models as a tool for investigating compositional effects, serious problems can also be identified: One of the main limitations is that they typically ignore potential unreliability in educational data: they assume that measurement is error-free; this is overoptimistic bearing in mind the fact that measurement error is the rule rather than the exception in educational research.

1.3 Unreliability in Educational data due to the Sampling of Items (Measurement Error) and due to the Sampling of People (Sampling Error)

Measurement error in educational data may manifest itself with different facets (see, for example, Webb and Shavelson, 2005). When students' achievement is assessed using item-level tests (as, for example, in large-scale international assessments and in standardized tests), measurement error is the result of the unreliability of the set of indicators used to assess a students' ability: the finite number of items that are used in the construction of tests are just a subset of the potential infinite items that could have been used to provide a perfectly reliable measurement.

In particular, in compositional models that involve higher-level aggregates of individual-level variables, measurement error can also occur at level 2: group specific influences may also distort the measurement of the intended level 2 construct (Lüdtke et al. 2008). For instance, situation specific factors in the way tests are administered in a certain class or school (e.g. a very strict teacher supervising the students or a fire alarm going off during the testing) or specifics of the item content (e.g. students from schools in different areas conceptualizing differently the notion of a set of questions) or, even, when a different person marks tests coming from the same school or class, are all factors that can result in systematic error variance at the group (school or class) level (see also Woodhouse, Yang, Goldstein and Rasbash, 1996).

Another source of error can be observed when individual-level variables are aggregated to obtain the group-level constructs – in addition to measurement error: this is error due to sampling. Whenever just a sub-sample of what is considered the total population of individuals for a cluster is used to obtain the higher-level unit, then the derived score is just an approximation of the true value of the group-level construct. Sampling error may occur in aggregated data in two different ways: when only a sub-sample from a finite cluster population is used in the analysis; or, when the assumed underlying population of the cluster is (potentially)

infinite. In the latter case, the sampling error is prevalent even if all individuals are sampled from within each cluster. Whenever I mention sampling error without any further comment in my thesis, I refer to the error arising under the assumption of infinite cluster population.

1.4 Integrating Multilevel and Structural Equation Models in a Single Analytical Framework

The methodological framework typically being used until recently to control for unreliability due to measurement error was one of single-level confirmatory factor analysis and Structural Equation Modelling (SEM; see Marsh, 2007a). In the SEM framework, the observed values of the variables involved in the modelling are conceptualised as realisations of the variables' true score (that is, a latent variable itself) probably measured with error. SEM research is concerned with issues related to the factor structure: how multiple indicators are related to the latent variables (factors) that they are intended to represent, the assessment of measurement error and the investigation of relationships among the latent variables, after controlling for measurement error (Marsh, Lüdtke et al., 2009). The problem with using these models in educational settings is that, conventionally, they fail to take potential clustering in the data into account. Moreover, the issue of sampling error in group aggregates is not easily addressed when using the SEM framework.

Only recently have these two dominant approaches in educational research, multilevel modelling (see section 1.2) and structural equation modelling, been integrated into a single framework. Using the Big-Fish-Little-Pond-Effect hypothesis as their substantive basis - a classic compositional effect widely investigated in the field of educational psychology (see section 1.8; also see section 2.1 in the literature review). Marsh, Lüdtke et al. (2009) demonstrated a 2x2 taxonomy of multilevel structural equation models. The set of models that they proposed can accommodate both nested structures in the data, and unreliability due to measurement error. Moreover, they allow adjustments for sampling error in the data (Lüdtke et al., 2008). Marsh, Lütke et al., used the term "manifest" in relation to measurement error or

sampling error when no adjustments are made for the corresponding source of error, and "latent" when measurement or sampling error is adjusted for. In this way, the doubly manifest model is the conventional compositional model that makes no adjustments for measurement or sampling error while the doubly latent model accommodates both measurement error at level 1 and level 2 as well as for sampling error in the higher-level aggregates. The models control for measurement error using multiple indicators and for sampling error assuming latent rather than manifest aggregation.

The methodology proposed by Marsh, Lüdtke et al. (2009) can be relevant for any researcher concerned about how individuals may be affected by their interaction with other individuals within similar settings and whenever the variables involved in the analysis are subject to unreliability (e.g. in econometrics and health, in organisational psychology; see Bliese, 2000 or in sociology; see Iverson, 1991 and sub-disciplines of psychology; Croon and Van Veldhoven, 2007). Specifically, in relation to school (and teacher) effectiveness studies, these methodological advances can be applied whenever the substantive interest lies in the investigation of the impact that the characteristics of the fellow students in a school (or a class) have on an individual's outcomes. In the following section I explain how I use the Marsh, Lüdtke et al. (2009) methodological framework to address ongoing debates on school compositional effects.

1.5 Substantive Focus: School Compositional Effects of Mathematics Achievement

With educational effectiveness studies, compositional effects of classroom-level and/or school-level aggregates of students' characteristics (e.g. achievement or socio-economic status) are often used to assess the effects of school composition (or school context; see section 1.1 on terminology) on students' outcomes (see Creemers Kyriakides and Sammons, 2010). School composition refers to the collective characteristics of the students in the school; the effects of the school composition then reflect the impact of between-school differences in their student

intakes on an individual's performance. School compositional effects denote the impact of a specific variable (such as the students' socio-economic status or achievement) as an aggregated variable at the school-level on students' outcomes over and above the contribution of the same variable in the model at an individual-level (see Harker, 2004). A significant positive compositional effect of, say, school average achievement suggests that students of the same academic achievement will benefit more academically if they go to an institution with higher achievement students. On the other hand, a negative compositional effect of achievement would suggest that attending a school with higher achievement might actually lead to lower subsequent outcomes as compared to attending a school with lower average achievement. It should be noted that such relationships are correlational and no causal inferences can be made.

Mathematics is one of the main subjects of the national curriculum in England and around the world. Achievement in Mathematics is not only an important educational goal in itself but it also has numerous applications in areas such as engineering, physics, economics and mechanics. It would not be an exaggeration to claim that Mathematics is the basis of all sciences. Interestingly enough, Reyna and Brained (2007) claim that a good knowledge of mathematics can even play an important role in making accurate health and social judgements in everyday life. Moreover, Osmond (2000) warns employers that a person's mathematical ability is an important predictor of his/her ability to carry out effective implementation of everyday work tasks. Because of the value of mathematics as an instructional goal, I consider mathematics-related outcomes for the purposes of my thesis.

My focus is on school average prior mathematics achievement and the use of this construct to quantify the effects of school composition; that is, I focus on school compositional effects of prior achievement in mathematics. Although other measures have been incorporated to operationalize the effect of school composition in the literature, including average socio-economic status (Willms, 1986), gender composition (De Fraine, Van Damme, Van Landeghen, Opdenakker, and Onghena, 2003, Kyriakides and Tsangaridou, 2008) and even the ethnic composition of the school (Duru-Bellat, Bastard-Landrier, Piquée and Suchaut, 2004), I choose

to focus on measures of achievement. An uncontroversial finding from the existing literature is that prior achievement at the individual-level is by far the best predictor of future outcomes (Thrupp, Lauder and Robinson, 2002). Thus, the extent to which the same variable at the aggregate level is an equally important predictor in explaining differences in the outcomes of students is investigated.

1.6 The Importance of Correct Estimation of School Compositional Effects

The substantive importance of the enquiry into whether or not school compositional effects exist is closely linked to the question whether "Schools Matter" (Mortimore, Sammons, Stoll, Lewis and Ecob, 1988). For instance, if the school composition has no impact on student achievement, then school effectiveness varies only as a function of the instructional and management practices of the school; thus, the school is highly accountable for its students' performance (Thrupp et al., 2002). On the other hand, if compositional effects are substantial, then school effectiveness may be limited by the nature of the school intake and the way in which students interact with each other in that social context. In this second scenario, it is not clear how responsible schools actually are in raising their students' performance.

Investigating school compositional effects is also very important at a political level – for the development of effective educational policies – and at a personal level – for parents seeking to choose a school for their children (see also Willms, 1985a). Relevant research may prove to be valuable in addressing the impact of social segregation or policies such as parental choice, selective schooling, the de-centralising of the school system and the identification of each school as an autonomous organization – practices that have been identified as leading to the polarization of intakes (see Thrupp, 1999; Gibbons, Machin and Silve, 2008). If schools are found to be effective irrespective of their intakes (i.e. no significant effects of school average ability on subsequent student outcomes), then this suggests that existing social norms underlying an educational system or the practice of educational marketization do not really lead to a disproportionate transition of knowledge. On the contrary, if school effectiveness is indeed

found to be linked to the school composition (i.e. a significant effect of school average prior achievement), then that should be a matter of concern to the educational community. It would suggest that the uneven distribution of intakes results in an unfair provision of knowledge among schools (Dale, 1977; see also Thrupp, 1999). In this way, evidence for or against the prevalence of school compositional effects relates to the question whether or not schools can compensate for society providing equal opportunities for all irrespective of their background and the initial academic knowledge that this background suggests.

It should now be clear that school compositional effects play a central role in the field of school (and teacher) effectiveness studies (see also Ballou, Sanders and Wright, 2004; Guldemond and Bosker, 2009; Verachtert, Van Damme, Onghena and Ghesquière, 2009; Van de Grift, 2009). Despite this, the state-of-the-art Marsh, Lüdtke et al. (2009) multilevel structural equation models have not yet been routinely applied in relevant research. This is ironic since one of the main debates that pervade the assessment of the magnitude and direction of the effects of school (or classroom) composition is the bias in estimation due to mis-measurement of relevant variables (Dumay and Dupriez, 2008; Harker and Tymms, 2004; Hutchinson, 2004; see also section 1.7; see further section 2.3.8 in the literature review chapter). A key purpose of my thesis is to adapt these methodologically stronger models from academic self-concept research into school effectiveness research, particularly in relation to the more appropriate estimation and interpretation of the effects of school-level average achievement, in the way this construct is used to index school composition.

1.7 Addressing the Issue of Under-Specification at Level 1 due to Mis-measurement of Relevant Variables

Research into the effects of school composition has been prevalent since the emergence of educational effectiveness paradigm (see Coleman et al., 1966). However, empirical studies on school compositional effects have not yet reached a consensus on their magnitude and direction. Reference to relevant on-going debates can be found in Wilkinson et al. (2000), Lauder,

Kounali, Robinson, Goldstein and Thrupp (2007) and Thrupp, (1999; see also et al., 2002). Several reasons have been identified as contributing to this lack of consensus (see, for example, Wilkinson et al., 2000).

Methodologically, these ongoing debates can be, to an important extent, attributed to under-specification at level 1 in compositional models (see Televantou, Kyriakides, Marsh, Nagengast, Fletcher and Malmberg, in press). Under-specification at level 1 may be evident when an insufficient number of level 1 covariates are controlled for in the models (omitted variable bias) or when the level 1 indicator on which aggregation is based is mis-measured (measurement error bias). Both of these facets of under-specification at level 1 in compositional effects models have been shown to bias compositional effect estimates upwards or downwards (see Harker and Tymms, 2004; Burstein, 1980), leading to misleading effects of the aggregated variables—artefacts of the statistical procedures. In my thesis, I address the issue of unreliability due to measurement error in the variable on which aggregation is based.

In a study especially relevant to the present thesis, Harker and Tymms (2004) use data from the Performance Indicators in Primary School (PIPS project; see Tymms, 1999) and show how a very significant positive school compositional effect of prior achievement can appear, simply by reducing the reliability of the pupil-level measure on which it is based. The researchers systematically added random error to pre-test measures that were highly reliable. The more measurement error they added, the larger the apparent positive effect of school average achievement became; schools with initially more able students were seen to be more effective. These effects were characterised as "phantom", simply an artefact of the failure to control for measurement error. Harker and Tymms explain that as the individual-level measure becomes less reliable it explains less of the variance. However, the aggregated measure does not lose its reliability to the same extent, since it is an average across a group of pupils. The aggregate measure is then well placed to explain variance at the school-level that should be attributed to the level 1 variable. Importantly, whilst Harker and Tymms clearly demonstrate the phantom effect and how it biases estimates of compositional effects, they do not propose any

statistical models to control for this bias since such models were not readily available when their study was conducted; this is a central focus of the present thesis: I use the Marsh, Lüdtke et al. (2009) methodological framework to address bias in the estimation of the effects of school composition due to mis-measurement of the variables involved in the compositional effect analysis (for more details on this methodological framework see section 2.3.8 of the literature review chapter). I build on previous studies that also have been concerned with spurious compositional effects that can arise due to poor measurement of the intake variables (e.g. Hutchinson, 2007; Ferrão and Goldsein, 2008; Gray, Jesson and Sime, 1990; see also Thomas and Mortimore, 1994; Woodhouse et al., 1996), inasmuch as I also consider unreliability in the aggregated measures both in terms of sampling of items (measurement error) and sampling of people (sampling error).

1.8 Considering both Cognitive and Affective Educational Outcomes

Most research on school effectiveness has focused solely on standardized test scores as the only outcome measure (Luyten and Veldkamp, 2011); there is little emphasis on other indicators of effectiveness: tests of the construct validity of test scores as a measure of effectiveness in relation to other indicators, or formative evaluations to improve effectiveness (Marsh, Nagengast, Fletcher and Televantou, 2011). Nevertheless, concerns expressed by researchers such as Sammons (1996) and Teddlie and Reynolds (2000) claim that research should be based on more than one effectiveness criterion: both cognitive and non-cognitive outcomes of schooling should be considered (see also Thomas, Sammons, Mortimore and Smees, 1997; Van de Gaer, De Fraine, Pusjens, Van Damme, De Munter and Onghena, 2009; Kyriakides and Luyten, 2009). Emphasising this issue, Goe, Bell and Little (2008, p. 52) state that: "There is no single measure that captures everything that a teacher contributes to educational, social, and behavioural growth of students, not to mention ways teachers impact on classrooms, colleagues, schools and communities." Therefore, there is a need to apply more sophisticated models of compositional effects to investigate a broader range of critical educational outcomes thus

providing a more inclusive evaluation of educational effectiveness. In this respect, investigating school compositional effects of prior mathematics achievement on students' affective outcomes is just as important as investigating these effects in relation to subsequent mathematics achievement.

In my thesis, I focus on academic self-concept with respect to mathematics. Academic self-concept is an important educational outcome in its own right (see also Nagengast and Marsh, 2011, Xu, 2010) and it also predicts other desirable outcomes such as academic choice, subsequent academic achievement, educational aspirations, and long term engagement (Guay, Marsh and Boivin, 2003; Marsh, 1991; Marsh and Craven, 2006). Indeed, there is an increasing body of evidence showing that academic self-concept is a critical outcome variable that in some cases is even more important than standardized test scores in predicting long-term achievement and critical academic choices (e.g., Marsh, 2007b).

Previous research has consistently found a negative compositional effect of achievement on academic self-concept, despite the fact that individual achievement is positively related with a students' academic self-concept (Marsh and Parker, 1984); this is the well-known Big-Fish-Little-Pond-Effect (BFLPE). The negative effect of school- and class- average ability on academic self-concept generalizes well over different countries and different levels of education (see Marsh, Seaton et al., 2008). An important contribution of my thesis is to demonstrate this effect for younger students in the first four years of their education.

School compositional effects of school average ability on self-concept and the BFLPE were the basis on which the 2x2 taxonomy of models was developed; several empirical studies have used this framework to address big-fish-little-pond-effects since the initial demonstration of Marsh, Lüdtke et al. (2009), for example, Nagengast and Marsh (2011; 2012; see also Parker, Marsh, Lüdtke and Trautwein, in press). Still, my thesis is apparently the first time that doubly latent models are applied to investigate compositional effects of achievement on self-concept for students as young as five to seven years of age.

1.9 Bridging the Gap between Self-Concept Research on Big-Fish-Little-Pond-Effects and Educational Effectiveness Studies on School Compositional Effects

The negative compositional effect of school average achievement on students' self-concept, as suggested by the BFLPE hypothesis, contradicts the assumption made by many parents in choosing a school for their children, namely that the higher the achievement of the intake of the school, the better the quality of education that it provides. It is also at odds with empirical findings reported by that part of educational effectiveness research that claims that students attending a higher achievement institution are likely to perform better than they would in a school with the opposite composition of students (see, for example, Lauder et al., 1999; Summers and Wolfe, 1977). It should be noted that some of the positive compositional effects of prior achievement reported in relevant educational effectiveness research are weak (e.g. Smith and Tomlinson, 1989; Gray et al., 1990) and even negligible (Thomas and Mortimore, 1994). In addition, the conceptual and methodological problems associated with theorising and modelling such effects (see, for example Thrupp et al., 2002) allow no space for valid inferences based on these findings alone.

In contrast, the findings on the Big-Fish-Little-Pond-Effect are well-grounded in a rich theoretical literature drawing from educational psychology, social psychology, psychophysical research, and sociology. There is also solid evidence that academic self-concept and academic achievement are reciprocally related (Marsh, Craven and Debus, 1998; Marsh, Trautwein, Lüdtke, Köller and Baumert, 2006). Furthermore, there is at least some academic self-concept research that suggests that school-level achievement has negative effects on other variables – including subsequent achievement -- that is mediated at least in part by academic self-concept.

In my thesis, I will show that the apparently positive effects of school average achievement on subsequent individual-level achievement is at least partly due to sub-optimal statistical models upon which this research has been based and continues to use. In this respect I

hope to bridge the apparent gap in academic self-concept research findings of negative effects of school average achievement and educational effectiveness research finding of positive effects of school average achievement.

This integration of two areas of research culminates in a longitudinal study of compositional effects in primary schools that looks at the effects of school average achievement separately for students' subsequent achievement and academic self-concept, integrating these two outcomes into a combined model of compositional effects on both cognitive and non-cognitive outcomes (see Study 1 and Study 2 of my thesis and their description in section 1.11).

1.10 Using school aggregates of achievement to explain between school differences in their absolute effects

In the compositional models that are used in my thesis to assess the effect of school composition on students' mathematics achievement, school-level aggregates of prior achievement are used to explain differences between schools in their students' achievement outcomes (achievement at the second measurement point) after adjustments for achievement at the first point (Sammons and Bakkum, 2011). In this respect, the compositional models that I incorporate resemble value added models (see section 2.3.5) that solely make adjustments for prior achievement and school average prior achievement. With multilevel value added models of educational effectiveness the percentage of residual variance in students' achievement scores at the level of the school is used to gauge school effectiveness – this is used as a measure of relative school effects. When school compositional effects are incorporated into value added modelling, they are then conceptualised as the effects of school-level factors that may potentially influence achievement but over which the school has no control at all. These are often distinguished from the effects of the school practices and policies (see, for instance, the discussion on "Type A" and "Type B" school effects in Raudenbush and Willms, 1995; see also section 2.3.5).

While value added models of educational effectiveness have comprised the conventional approach to school accountability and the basis for the construction of the schools'

league tables (Van de Grift, 2009), an alternative approach to studying school effectiveness has recently been proposed (Cook, 2008; Kyriakides and Luyten, 2009; Luyten, 2006; Luyten, Peschar and Coe, 2008; Luyten, Tymms and Jones, 2009): this is the regression discontinuity approach. Once the main assumptions underlying this approach are fulfilled (see section 2.5.5), it provides estimates of added-year effects, that is, of the effect that one extra year of schooling has on students' academic achievement. When implemented within the multilevel modelling framework it can also provide measures of the extent to which schools differ in their absolute schooling effects; thereby providing relative measures of effectiveness in addition to absolute measures of effectiveness.

An important advantage of the regression discontinuity approach compared to the multilevel value added models is that, in principle it does not require controlling for prior achievement or any other background characteristics in order to assess the effect of schooling. Importantly, individual achievement is used in regression discontinuity models as a dependent rather than a control variable. This is of considerable relevance to the two sources of under-specification that have been discussed in section 1.7 omitting or mis-measuring relevant covariates. While these can be especially problematic in compositional effects models and value added models of educational effectiveness, regression discontinuity estimates should be no more prone to the issue of bias arising due to these sources of error.

Given this strength of the regression discontinuity design as compared to conventional multilevel models (and also other advantages of the regression discontinuity models as compared to compositional effects models and value added models of educational effectiveness in section 2.5.6), it is important for educational researchers to further explore the possibility of using this methodological tool for purposes of school accountability. In response to this need, I incorporate regression discontinuity models in my thesis in order to investigate potential effects of school composition on added-year effects by including school-level covariates as predictors in the slope of the absolute effect of schooling: In the same way in which school-level aggregates of student-level characteristics can be used with value added models to explain

variability in relative school effects, they can be used with regression discontinuity approaches to explain variability across schools in their absolute effects (see, for example, Luyten, 2006; Luyten, Peschar and Coe, 2008). The question is then whether the effect of one extra year of schooling correlates with school composition.

In contrast with research into the use of school-level aggregates aiming to capture the effects of school composition in compositional effects models and value added models of educational effectiveness, the existence of measurement error in students' achievement has not been of major concern in previous regression discontinuity studies (e.g. Cliffordson and Gustafsson, 2011; Luyten, 2006; Luyten et al., 2008; 2009). Perhaps this is because the impact of measurement error on the response variable in regression models is known to have little or no impact on the estimated coefficients per se (e.g. Woodhouse et al., 1996); it only leads to larger standard errors that reflect the increased unreliability in the estimation (the distribution of the estimated effects have larger standard deviations). Nevertheless, considering that measurement error in the criterion can result in a decrease of the power of the statistical analysis (the probability of detecting a significant effect when this effect holds place in the population), there arises the need to be able to correct for measurement error in the regression discontinuity models and retrieve the correct standard error estimates. Indeed, one limitation that the regression discontinuity approach has been accused of relative to the random assignment design – for which it is supposed to serve as an alternative (see section 2.5.6) - is that much larger sample sizes are required to achieve estimates with the same level of statistical power (Schochet, 2009). This has been demonstrated, for example, by Goldberger (1972) who showed that for a non-clustered design the sample under the regression discontinuity design must be 2.75 times larger than for a corresponding experiment to achieve the same level of statistical precision. Moreover, the school sample sizes typically need to be about three to four times larger under regression discontinuity than true random assignment designs to achieve impact estimates with the same levels of precision (Schochet, 2009). Considering the small sample sizes of some of the studies that have been conducted within the field of educational

effectiveness (e.g. Luyten et al., 2009), not adjusting for measurement error – which has an additional detrimental effect on the power of the study – may be considered as a potentially important limitation in the application of the regression discontinuity approach within the school effectiveness paradigm.

The present study evaluates the impact of measurement error on the parameter estimates and the associated standard errors of the regression discontinuity designs. In light of new methodological developments (section 2.2.5) that allow corrections for measurement error through the use of multiple indicators, it shows how higher accuracy in estimation can be achieved. Crucially, by integrating the regression discontinuity approach with multilevel structural equation models, it is demonstrated how it is possible to estimate the impact of school-level aggregates of student-level characteristics on added-year effects, controlling for both measurement error and sampling error bias.

1.11 An Overarching Organization of the Thesis

The thesis is comprised of five chapters: Chapter One is the present chapter, i.e. the Introduction, and Chapter Two reviews relevant literature. In Chapters Three , Four and Five, I present the three main studies of my thesis, respectively: "Phantom Effects in School Composition and Self-Concept Research: Consequences of Failure to Control for Bias due to Measurement Error in Traditional Multilevel Models" (Study 1), "The Big-Fish-Little-Pond-Effect: Evidence from Early English Primary School Data" (Study 2) and "An application of the Regression Discontinuity Approach to English TIMMS 95 data: Addressing the Issue of Measurement Error Bias and Exploring the Possibility to Investigate Potential Effects of School Composition" (Study 3). In the presentation of each study I begin with an introduction to the main aims of the analysis undertaken. I then present the research questions and research hypothesis underlying the study. The methodology followed is subsequently outlined and relevant results are summarised. The findings of the three studies come together with a comprehensive discussion in the concluding chapter, Chapter Six. In the discussion chapter, I

explain how the methodology employed in each can be a promising methodological tool for educational researchers seeking to investigate the potential effects of school composition on students' cognitive and affective outcomes. In an Appendix I include (i) some more technical parts relevant to the topics addressed, for example relevant mathematical equations and derivations of mathematical formulas (ii) supplementary analyses to the statistical analyses undertaken (iii) further information on topics raised in the literature review on which I felt I should expand somewhat more than I had already done in the main thesis document (iv) useful Mplus syntax.

I begin my literature review (Chapter Two) with the Big-Fish-Little-Pond-Effect - this is the historical basis of the methodological framework that I use throughout my thesis. I review the range of methodological approaches that have at times been used for the investigation of this phenomenon in self-concept research following its very first demonstration (Marsh and Parker, 1984). I explain that the Marsh, Lüdtke et al. (2009) 2x2 taxonomy of models are the most recent methodological advances ever applied to address big-fish-little-pond-effects.

In a separate section I review the Marsh, Lüdtke et al. (2009) framework, highlighting its advantages as compared to conventional compositional models. I begin with a reference to the approach conventionally used for the potential effects of the school composition. I then discuss the consequences of not adjusting for measurement and for sampling error in this model. The four models of the 2x2 taxonomy are introduced, explaining how each addresses different sources of unreliability prevalent in the educational data. I explain the usefulness of partial and doubly latent models for educational effectiveness research and in particular for the investigation of school compositional effects. In this way, I provide a linkage with the third section of my literature review which is dedicated to the main substantive focus underlying my thesis, namely the school composition effect.

In the third section of my literature review, which reviews studies on the potential effect of school composition, I note the difficulty of disentangling this effect from the effect of the

schools' policies. Theories on how the school composition can influence students' outcomes are summarized and the debate around the size and the direction of school compositional effects is outlined. Importantly, I explain the methodological flows underpinning the estimation of such effects explaining how the present thesis addresses them.

Lastly, the fourth section of the literature review is dedicated to the regression discontinuity approach. The need to assess the absolute effect of schooling and the difficulty in doing so are discussed. The use of the regression discontinuity approach as a statistical tool to investigate absolute schooling effects is then outlined and the assumptions underpinning the approach are explained. The chapter ends with a review of recent studies that have used the approach to investigate the relationship between school composition and added-year effects.

In Chapter Three, I present my study on the application of the Marsh, Lüdtke et al. (2009) 2x2 taxonomy of compositional effects models to investigate school compositional effects of school-average prior mathematics achievement on students' subsequent mathematics achievement. Study 1 consists of two distinct but inter-related studies:

Study 1a: Using English primary school data from the Performance Indicators at Primary School Project (PIPS), I replicate Harker and Tymms's (2004) results in their classic demonstration of phantom compositional effects. The focus is on the compositional effect of year one mathematics achievement on students' year four mathematics achievement. As a first step, I add random noise to the student-level baseline measures (year one mathematics achievement) producing achievement scores with lower reliability and observing the impact that this has on compositional effects estimates. I extend the Harker and Tymms demonstration, showing that the full correction and partial correction approaches of the 2x2 taxonomy are able to compensate for measurement error even for data that have large amounts of measurement error.

Study 1b: In this study, I use data on the mathematics achievement at the beginning and at the end of primary grade[1] four for a sample of Cypriot students. The focus is on the school compositional effect of prior mathematics achievement on students' subsequent mathematics achievement and the extent to which this effect is prone to bias due to measurement and/or sampling error. Study 1b is of particular importance, especially for the purposes of the main project for which the database incorporated has been originally used (the "Promoting Quality in Education" project). The main aim of the "Promoting Quality in Education" project has been on the role of factors operating at different levels (student, teacher, school and classroom) on students' achievement. Correct estimation of compositional effects at the level of the school, the focus of Study 1b, is crucial in assessing the impact of school-level factors on students' outcomes and especially in determining its relative importance to the effect of the school's policies and practices (e.g. the school's policy on teaching and the school learning environment). Moreover, this is the first time the models of the 2x2 taxonomy are applied to investigate compositional effects using primary school data from Cyprus. While the educational systems of England and Cyprus are in some ways similar, there are also some notable differences between them. Hence, juxtaposing the results of the same school effectiveness model across data from the two countries can provide insights into the generalizability of the findings in different educational contexts.

In Chapter Four, I present the second major study of my thesis. The analysis conducted as part of Study 2 is parallel to that in Study 1a, in that it uses the same dataset (PIPS) and in that they both consider the compositional effect of school average mathematics achievement at year one on students' subsequent outcomes. Here, however, I investigate school compositional effects of mathematics achievement on a different educational outcome; namely the students'

[1] Note here the use of grade rather than year. In Cyprus, distinct phases of schooling are referred to as grades while in England they are most often referred to as years.

mathematics self-concept, verifying the big-fish-little-pond-effect hypothesis. Doubly latent models have been the latest methodological development in investigating big-fish-little-pond-effects (see Seaton and Marsh, 2012); these cutting edge approaches are incorporated for the purposes of Study 2. All four models of the 2x2 taxonomy are applied to the data, and the resulting estimates are compared. Hence, the importance of correcting for measurement and sampling error in big-fish-little-pond effect models is gauged.

In two separate set of analyses, I consider the compositional effect of (i) school average mathematics achievement in year one on students' self-concept at the end of year one and (ii) school-average achievement in year one on students' self-concept at the end of year four. Big-fish-little-pond-effects across the distinct phases of schooling are compared to each other. Lastly, bridging the findings of Study 2 with the findings of Study 1a, I examine the extent to which academic self-concept at year four mediates the relationship between school average achievement in year one and individual achievement in year four.

In Chapter Five I present the third study of my thesis (Study 3). In Study 3, I use English mathematics achievement data from the Third International Mathematics and Science Study conducted in 1995 (TIMSS 95) and I apply the regression discontinuity approach to investigate the absolute schooling effect at these two distinct phases of schooling as well as differences between schools in bringing about this effect. With multilevel regression discontinuity models the measure of effectiveness is the variability of the effect of one extra year of schooling across schools.

In two parallel analyses, one for primary school data (years four and five) and one for secondary school (years eight and nine), I investigate a set of research questions and research hypotheses related to the robustness of the regression discontinuity estimates (i) to adjustments for different student background characteristics and (ii) to adjustments for measurement error in students' achievement scores; this is used as the criterion in regression discontinuity models in their applications to measure educational effectiveness.

Methodologically this study contributes to existing knowledge in that it integrates the regression discontinuity approach with multilevel structural equation models to control for unreliability in students' measured academic achievement. Apparently, it is the first study to use the achievement data released in April 1999 that contain scaled scores not only for the overall mathematics achievement of students but also for their achievement in the different content areas (Whole numbers, Fractions and Proportionality, Geometry and Measurement, Data Representation, Analysis and Probability). The latter are used as multiple indicators to control for measurement error in the measured achievement of students.

Perhaps the most important component of Study 3 from a substantive perspective is to evaluate the extent to which variables related to the composition of the school (school average mathematics achievement and the proportion of students' in the school that come from disadvantaged economic backgrounds) explain a significant proportion of variability across schools. In this way I demonstrate how regression discontinuity models can serve as an alternative to compositional effects models and value added models of educational effectiveness in the estimation of the effect of school composition. Importantly, I demonstrate how relationships between school composition variables and added-year effects can be evaluated in the multilevel structural equation modelling framework so that (i) measurement error at level 1 and level 2 and (ii) sampling error in school-level aggregates are adjusted for.

In Chapter Six I discuss the findings of my studies (Study 1, Study 2 and Study 3) simultaneously, pointing out the strengths and weaknesses of each: that is, the way in which they add to existing literature and their potential limitations, and highlighting issues that would merit further investigation.

Chapter 2: A Review of the Literature

The aim of this chapter is to provide evidence of extensive reading on the general and specific topics addressed in the present investigation. To be precise, the chapter is organised into five sections. I begin with the literature on the Big-Fish-Little-Pond-Effect, stressing the relevance of the present investigation to both substantive and, importantly, methodological issues that pertain to the field. I then explain how methodological advances stemming from self-concept and Big-Fish-Little-Pond-Effect research can be used to address the issue of measurement and sampling error bias in any analyses concerned with the estimation of compositional effects. Indeed, these models can be incorporated for the assessment of school compositional effects in educational effectiveness research. I discuss this in detail in the third section of my literature review, explaining how the use of this methodology can contribute to relevant on-going debates. The fourth section is dedicated on the use of the regression discontinuity approach in educational effectiveness research – the focus of the third study of my thesis.

2.1 The Big-Fish-Little-Pond-Effect: Theoretical Basis, Generalizability across Age and Different Educational Outcomes, Measurement Issues and Methodological Advances

The Big-Fish-Little-Pond-Effect (BFLPE) is a major focus of my thesis (Study 2 of my thesis). More importantly, it has stimulated the development of the statistical tools that I incorporate in order to address biases in school compositional effects due to (multilevel) measurement error (Study 1 of my thesis).

I begin this first section of my literature review with a description of the BFLPE paradigm, reviewing relevant literature and highlighting its wide generalizability across different research settings. Importantly, I outline studies that have been conducted since the very first demonstrations of this phenomenon (see, for example, Marsh, 1987; Marsh and Parker,

1984) and I explain how the methodology incorporated to address the BFLPE has been changed over time to address measurement and statistical issues that pertain to it. The theory of self-concept is the theoretical basis of the BFLPE model. Hence I begin with a brief summary of academic self-concept theory and research, before moving on to BFLPE.

2.1.1 Self-Concept

In defining self-concept in a broad way, Shavelson, Hubner and Stanton (1976) designate it as a person's perception of self. They explain that these perceptions are formed through interaction with one's environment and are particularly affected by environmental influences and significant others. To be precise, the individual's own reference group, defined generally as the group to which a person belongs or aspires, has been recognized by researchers as an important factor to consider in understanding self-concept in the relevant area decades ago (see for instance, Purkey, 1970). This is especially relevant to the BFLPE (see section 2.1.3) which assumes that students form their self-concept (academic; see section 2.1.3 on the domain specificity of the BFLPE) by comparing their own accomplishments with those of their school -(or class -) mates.

Originally, self-concept was conceived as a unidimensional construct: the assumption was that there was only one general factor of self-concept or that a general factor dominated more specific factors (Coopersmith, 1967; Marx and Winne, 1978). Nevertheless, subsequent studies supported the construct validity of a multi-dimensional (Marsh, 1990; Marsh, Byrne and Shavelson, 1988, Marsh and MacDonald Holmes, 1990) and a hierarchical construct (see Marsh and Craven, 2006). The multi-dimensionality of self-concept implies that it consists of multiple components, all of which are correlated with each other – but still distinguishable from one another. The fact that self-concept is hierarchical in nature suggests that the lower-level, more specific components of self-concept, despite being distinct, share common variance that is attributed to an overarching higher-order self-concept factor.

One possible representation of self-concept is given in Figure 2.1. The overall higher-order factor, i.e. the general self-concept, is often defined as self-esteem, appraisal, worth (Marx and Winne, 1978; Rosenberg, 1965; Rosenberg and Simmons, 1971). Then in the lower-level of the hierarchy can be seen academic self-concept and non-academic self-concept (social self-concept, emotional self-concept and physical self-concept). These are further subdivided into sub-facets of self-concept that involve more specific aspects. For example, academic self-concept can be split into subject-specific components (English, Mathematics, Science etc). Shavelson et al. (1976) found modest support for this model. More recently, Marsh and Shavelson (1985) showed that it might be more appropriate to conceptualize academic self-concept as consisting of two second-order factors, one of verbal ability and one of mathematics. Thus, according to the Marsh/Shavelson model, instead of one second order factor corresponding to academic self-concept (see Figure 2.1), there should be two distinct factors at the same level, one for maths and one for verbal self-concepts.

Although not a main focus of the present investigation, the construct validity of a multidimensional perspective of self-concept is extremely important (see Marsh and Craven, 2006) in that the most powerful effects of self-concept (see for example the BFLPE and its domain specificity in section 2.1.3) are based on specific rather than global components of self-concept; these in turn are related to specific outcomes (a multidimensional rather than a uni-dimensional perspective). Relevant theoretical models (e.g. Marsh, 1991; Marsh and Craven, 2006) are important in tests of the BFLPE. In particular, in the context of education and BFLPE research (see section 2.1.3) important educational outcomes are systematically related to academic self-concept, but relatively unrelated to self-esteem.

Indeed, current conceptualizations of self-concept are domain specific. For example, mathematics self-concepts are more strongly associated with mathematics achievement than with achievement in another domain such as English or Science (e.g. Marsh, Trautwein, Lüdtke, Köller and Baumert, 2006). This domain specificity of self-concept provides support for the

importance of the multidimensionality of self-concept that, in turn, provides support for the construct validity for the interpretations of the BFLPE (see section 2.1.3).

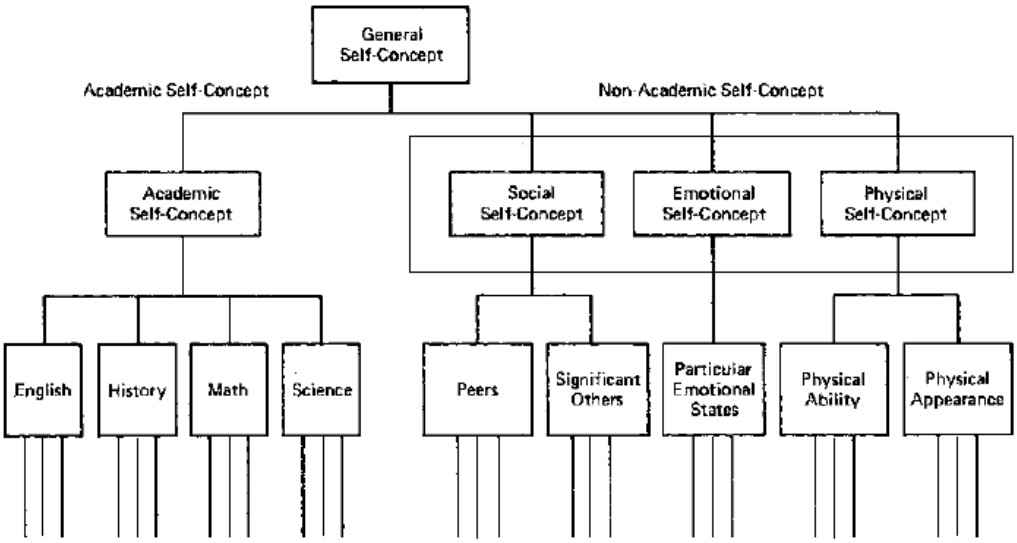
In what follows, I explain that while academic self-concept is an important educational outcome in its own right, it also facilitates other educational outcomes and should therefore also be considered in educational effectiveness research (see section 2.3.1) together with academic achievement.

2.1.2 Academic Self-Concept: An Important Educational Outcome in its own Right

In educational research academic achievement is the critical variable (Marsh and Craven, 2006). However, throughout the history of educational effectiveness (see section 2.3.1) educational goals have been fluctuating: while the initial focus was solely on cognitive outcomes (Callahan, 1962), educational research has also been concerned with social and affective ones. Recent studies emphasize the need to address the latter (Sammons, 1996; Teddlie and Reynolds, 2000). The enhancement of the child's self-concept in particular has been recognized as a major educational goal for over forty years now:

> It has become increasingly clear in the light of the schools' attempt to serve the disadvantaged that the schools have a fundamental responsibility to enhance the self-concepts of their students (Zirkel, 1971, p. 211)

Figure 2.1: One possible representation of the hierarchical organization of self-concept (Shavelson, et al., 1976, p. 413)



Note. Non-Academic components of self-concept according to the present diagram are Social Self-Concept, Emotional Self-Concept and Physical Self-Concept. These are separated from the Academic component of self-concept. The different domain–specific self-concepts of which academic and non-academic self-concept are composed, as shown in the diagram, can be further divided into more specific self-concept components. For instance, science self-concept may be further sub-divided into biology, physics and chemistry. These components could have been displayed at a lower-level in the diagram and could, in turn, be divided further into more specific components.

Academic Self-Concept (ASC) refers to that specific component of self-concept that denotes the way in which individuals perceive their academic abilities and competencies in a specific subject (see Figure 2.1; Byrne and Shavelson, 1986). It is clear then that self-concept is valued as an educational outcome in its own right. But even if its potential value as an educational outcome in itself is neglected, its potential for interpreting achievement outcomes cannot be ignored.

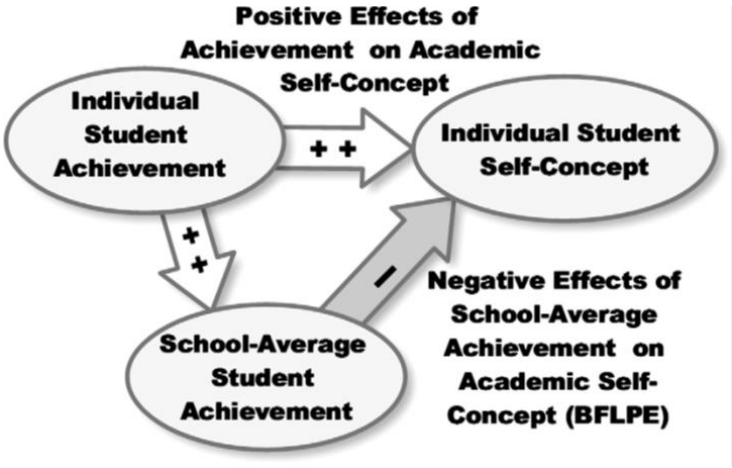
ASC acts as a mediator and helps in the development of other desirable psychological and behavioural outcomes, such as academic choice, educational aspirations and long-term engagement. It is a critical predictor for course and career choices, even more important than individual achievement when both are included as predictors in the same model (Guay, Larose and Boivin, 2004; Marsh and Yeung, 1997). Existing research (e.g. Marsh and Craven, 2006; Marsh and Yeung, 1997, Valentine and DuBois, 2005) has consistently shown mutually beneficial effects between academic achievement and academic self-concept (the reciprocal effects model, see for example Marsh, Craven and Debus, 1998; Marsh, Chanal and Sarrazin, 2006) with these relations being very domain specific and consistent over age (Marsh, 1992).

Positive self-beliefs are central in diverse fields other than education, fields such as child development, mental and physical health, social services, industry, and sport psychology. Recent developments in the positive psychology movement (e.g. Seligman and Csikszentmihalyi, 2000) imply that a positive self-concept is positively correlated with future life successes as well as reflecting previous life experiences (Marsh, Lüdtke et al., 2009).

Thus, academic self-concept, whether used as an outcome in itself or as a moderator variable that helps explain achievement outcomes, is a critical variable in education and in educational effectiveness research. A central focus of self-concept research is the "Big-Fish-Little-Pond-Effect" (BFLPE) hypothesis. This relates to educational processes taking place within classes or schools and their potential effects on the students' academic self-concept. I expand on this in the following section.

2.1.3 The Big-Fish-Little-Pond-Effect: A Definition

Figure 2.2: The Big-Fish-Little-Pond-Effect (Marsh, 2007b)



The Big-Fish-Little-Pond-Effect hypothesis (see Figure 2.2) predicts that individual student achievement (L1-ACH) has a positive effect (++ in Figure 2.2) on Academic Self Concept (L1-ASC) while the corresponding effect of the group average (school or classroom) achievement (L2-ACH) is negative (-) (Marsh, 1987; Marsh and Parker, 1984; Marsh, Seaton, Trautwein, Lüdtke, Hau, O' Mara and Craven, 2008). In this way, it suggests that students who have the same academic ability have lower ASC when they are in high ability classes or schools (i.e. classes or schools where the average achievement is high) than when they are educated in mixed or low-ability classes or schools.

The BFLPE is domain specific in that although this phenomenon has been widely demonstrated in relation to academic self-concept, the effects of school-average ability on global self-esteem or other non-academic components of self-concept are small or non-significant (Marsh, 1987; Marsh, Chessor, Craven and Roche, 1995; Marsh and Craven, 2006). Hence, the focus of the BFPE is on academic self-concept and on domain specific self-concepts such as Mathematics, English or Science.

2.1.4 A Theoretical Model of the BFLPE

The theoretical premise of the BFLPE emphasises that frames of reference must be considered to fully comprehend how people perceive themselves. Depending on the frames of reference or comparisons they use to evaluate themselves, individuals can reach different conclusions about their accomplishments and so have varying self-concepts. Although mainly applied in educational research, the historical and theoretical roots of the BFLPE frame-of-reference theoretical model come from a variety of disciplines including psychophysics (e.g., Helson, 1964; Marsh, 1974; Parducci, 1995; Wedell and Parducci, 2000), social judgment (e.g,. Morse and Gergen, 1970), sociology (e.g., Alwin and Otto, 1977), relative deprivation (e.g., Davis 1966) and social comparison (e.g. Festinger, 1954).

The theoretical model of the BFLPE as applied to academic self-concept in the educational psychology setting can be described by a series of predictions (see Marsh, 1984b), Support for such a model, based on psychophysical research, can be traced back to the 1990s (Mash, Seaton et al., 2008). Marsh (1984a) specifically developed the BFLPE paradigm to understand the formation of academic self-concept in school settings:

- ASC will be positively related to academic ability;
- School average ASC will be reasonably similar in high-ability and low-ability schools even though the ability levels of students can be substantially higher for the former and lower for the latter. This is a natural consequence of the fact that the frame of reference is mainly established by the student's own school;
- School average ability is negatively related to ASC after controlling for individual student ability;
- ASC will be more highly correlated with individual ability after controlling for school average ability;
- ASC is more accurately predicted from individual and school-average ability than from either of these covariates on its own

- The negative effect of school-average academic ability is mostly evident for ASC and not for other non-academic components of self-concept;
- The frame-of-reference is established by school-average ability; therefore all students in a high-ability school are predicted to have lower ASCs than the same students if they attended a low-ability school; interactions between school average and individual ability on ASC are expected to be small or non-significant.

Later studies (Marsh, 1987, 1991; Marsh and Craven, 2002) expanded on this theoretical model: the negative BFLPE was conceptualized as the net effect between a positive assimilation effect and a negative contrast effect (see also Marsh, Kong and Hau, 2000). Marsh (1987) explains that being an average ability student in a higher ability school may affect academic self-concept in three different ways: (i) negatively, with the basis of comparison being the performance of above average students (the BFLPE – a contrast effect), (ii) positively, because of the feeling of virtue being a member of a higher ability group (a reflected glory, group identification or assimilation effect), or (iii) not at all, because academic self-concept is unaffected by the immediate context of other students or because (i) and (ii) cancel each other out. Support for the BFLPE suggests a larger contrast effect but it does not imply that there are no assimilation effects that are overwhelmed by substantially larger contrast effects. However, empirical support for assimilation effects is weak (Marsh, 2007b).

Seaton and Marsh (2012) explain the process underpinning the BFLPE effect in a very clear way: Suppose that there are two students with similar above average intelligence, one of whom attends an academically selective institution (in which the school-average ability is higher than the average of all the other schools) and the other attends a non-academically selective school, where the school average ability is similar to the average of all the other schools). The student in the average ability school will then perform extremely well compared to his/her classmates and hence will develop positive attitudes towards his/her abilities (being a big fish in a small pond). However, the student who attends the academically selective school

has an average performance in relation to his/her classmates and, in this way, does not feel as confident in relation to his/her own abilities (a little fish in a big pond). In this way, the frame of reference associated with attending an academically selective school has resulted in two students with essentially the same ability having formed different perceptions in relation to that ability: The student in the academically selective institution feels less competent than the student in the non-academically selective school.

Although this difference in self-concept among students with similar academic excellence may seem unrealistic to external observers – after all both of these children seemingly have the same academic ability -- it is in perfect agreement with the child's own phenomological experience in comparing his/her ability with that of the other students in the school. In this respect, the BFLPE is a specific example of more general frame-of-reference effects with the standard of comparison being the school (or class) average achievement (Marsh, 1984b).

What I would finally note is that while the BFLPE suggests that social comparison will have detrimental consequences for students in selective academic environments, it implies quite the opposite for students in schools where the average ability level is low. They can benefit from such systems because they are grouped with other less achieving students and thus their frame of reference contains, on average, less able peers (Nagengast and Marsh, 2011). In the same way, high attainment students will have higher academic self-concepts in mixed ability settings where they are among the most able, than in selective settings in which the students are very bright (Marsh et al., 1995).

2.1.5 Generalizability of the BFLPE over Time (i.e. Stability of the BFLPE) and across Different Age Groups

One of the main aims of the research strategy of the BFLPE program is to show generality and ubiquity of the BFLPE over gender (Marsh et al., 1995), ability levels (Marsh and Craven, 2002; Seaton, Marsh and Craven, 2010) and cultures (Marsh and Hau, 2003; Marsh et al., 2007; Seaton, 2007; Seaton et al., 2010). Indeed, what makes this "paradoxical" finding (Marsh, 1984b, p.165) even more interesting is the fact that it can be generalized across different research settings. Of special interest to the present investigation has been the generalizability of the BFLPE over time (i.e. the stability of the BFLPE) and across students of different age. I refer more extensively to literature relevant to each of these in the two subsections that follow.

2.1.5.1 Stability of the BFLPE over time

Previous literature suggests that the negative effect of school average achievement on students' self-concept remains stable and even increases over time for students who remain in the same school setting. For instance, Marsh, Köller and Beaumert (2001) using longitudinal data from large cohorts of seventh grade German students demonstrated that the negative BFLPE was more negative in West German schools than East German schools at the start of the reunification of East and West Germany. They attributed this finding to the fact that West German students attended schools that were highly stratified in relation to ability before and after reunification, whereas East German students did not attend selective schools until after the reunification. Marsh et al. (2003) using a large sample of secondary school students in Hong Kong – a country that witnesses a highly segregated educational system - demonstrated that there was a substantial negative effect of school-average ability prior to the start of high school on grade nine ability even after controlling for the substantial negative effects in earlier school years. Hence, in their study, Marsh et al. (2003) also demonstrated the growth of the BFLPE over time. The negative effects of selective high schools on self-concept have even been

demonstrated several years after graduation (Marsh, Trautwein, Lüdtke, Baumert and Köller, 2007). Note, however, that the studies reviewed here are concerned with data from secondary school (or from older students who have graduated), in contrast with my research that compares big-fish-little-pond-effects for students in years one to four of primary education (see section 4.4.1 in Study 2 where I describe the sample that I use to test the BFLPE hypothesis).

2.1.5.2 Generalizability of the BFLPE across students of different age

Although most studies on the BFLPE come from secondary or high schools (e.g. Marsh et al., 2001), there is also some evidence for primary schools. Marsh et al. (1995), in two separate studies, one concerned with primary school children in years three, four and five (mostly nine to eleven years of age) and another with students ranging in age from nine-year-olds to eleven-year-olds, considered the effects of participation in gifted and talented programs on students' academic self-concept. They demonstrated systematic declines in students' academic self-concept over time; these results were consistent over age. This study is especially relevant to mine in that it is also concerned with primary school students; albeit somewhat older in age (see section 4.4.1 on the measures and data samples used in Study 2).

Evidence for the prevalence of the BFLPE for students as young as seven years old (in year two of their primary education) has been reported by Tymms (2001) using data collected in 1998 from English pupils as part of the Performance Indicators at Primary School Monitoring Project; mathematics attainment and self-concept data from this project have also been used for the purposes of my thesis (see section for Study 1 in which I describe the data used). In this study, Tymms (2001) found empirical support for "the view that being an able pupil amongst able pupils results in a decreased self-concept" (Tymms, 2001, p. 176). The researcher found a negative effect of class average academic level to attitudes towards Mathematics, Reading and School. Nevertheless, the reported effects were relatively weak.

2.1.6 Generalizability of the BFLPE across Different Educational Outcomes

What makes the BFLPE especially important is the fact that it has been demonstrated across a wide range of other desirable educational outcomes – as well as self-concept. Particular emphasis in the existing literature has been placed on the role of academic self-concept as a mediator of the negative effects of school average achievement on subsequent educational outcomes. Relevant research can be traced back to Davis (1966) who found that equally able students had higher career aspirations when studying in colleges with lower average ability; Davis referred to this phenomenon as the "frog-pond" effect. The suggestion made in this study was that the negative effect of school average ability on career aspirations was partially mediated by students' self-evaluations and their Grade Point Average (GPA). However, the data available were not adequate to fully test these predictions. Also, even before the establishment of the BFLPE, Alwin and Otto (1977) demonstrated that school average achievements have a negative effect on educational and occupational aspirations calling for further research to identify intervening processes mediating these negative compositional effects.

In one of the very early studies on the BFLPE, Marsh (1987) reported that school-average ability had a negative effect on both academic self-concept and school grades (Grade Point Average; GPA). In his study, Marsh stressed that the negative effects of school average ability on GPA and on academic self-concept were mutually reinforcing. He based this claim on the fact that frame-of-reference effects on GPA had indirect effects on subsequent academic self-concept and, conversely, frame-of-reference effects on academic self-concept had indirect effects on subsequent GPA. Further, he conducted a longitudinal analysis which suggested that academic self-concept had a direct effect on subsequent school grades and that part of this effect was due to the BFLPE. Significantly, Marsh (1987) found that the school-average ability had a direct negative effect at time 1 (T1) on GPA and a negative effect on GPA at time 2 (T2). The latter was mediated by academic self-concept among other time 1 (T1) variables.

In a subsequent study in response to the study conducted by Alwin and Otto (1987) (see previous paragraph), Marsh (1991) used longitudinal data from the "High School and Beyond" study of US high school students to consider the effects of school average achievement on a range of academic outcomes (for instance, standardized test scores, self-concept, coursework selection, academic effort, school grades, educational and occupational aspirations and college attendance). The outcomes were assessed in year ten, year twelve and again two years after graduation from high school. The effects of school average achievement were found to be negative for almost all of the year ten, year twelve and post-secondary outcomes: fifteen out of the seventeen effects were significantly negative and two were non-significant. School average achievement negatively affected academic self-concept (the BFLPE), educational aspirations, general self-concept, advanced coursework selection, school grades, academic effort, subsequent standardized test scores, occupational aspirations and subsequent college attendance. The negative effects for educational aspirations were clearly obvious two years after graduation from high-school. Interestingly enough, Marsh found in his study that controlling for the negative effects of school average achievement on academic self-concept substantially reduced the size of negative effects on other outcomes, consistent with the proposal that the negative effects of school average ability were mediated by academic self-concept.

Expanding on these results, Marsh and O'Mara (2010) demonstrated that school average achievement had a negative effect on academic self-concept, educational and occupational aspirations, and school grades – these effects were evident even five years after students had graduated from high school. In relation to the effect of school average achievement on educational attainment, they found that although the direct effect was apparently positive, the indirect effect – mediated by academic self-concept - was negative and so the total effect was non-significant.

Trautwein, Lüdtke, Marsh, Köller, Baumert (2006) considered the effect of school average achievement on academic interest. They demonstrated negative effects; they also

demonstrated that academic self-concept almost completely mediated BFLPEs on interest. Furthermore, Xu (2010) demonstrated that the BFLPE generalized to a range of other psychosocial constructs including importance, effort persistence, rehearsal, elaboration, and control strategies.

Although BFLPE research has primarily been related to academic outcomes, the BFLPE has also been verified in the context of physical as well as academic ability. For example, Chanal, Marsh, Sarazzin and Bois (2005) found a negative effect of the average physical ability of the group with which an athlete is trained on gymnastics self-concept. Trautwein, Gerlach and Lüdtke (2008) demonstrated that class average physical ability had negative effects on physical self-concept and long-term physical ability.

To sum up, the range of outcome variables considered to evaluate predictions of the negative effects of school average achievement in the BFLPE paradigm often expand over attainment, educational and occupational aspirations and long-term educational attainment. These findings suggest that, despite the fact that the theoretical foundation and implications of BFLPE have focused mainly on academic self-concept as the outcome variable, the policy implications of the BFLPE may be relevant not only to research on academic self-concept but also to other aspects of life potential.

2.1.7 Minimal Conditions for Testing the BFLPE

The Big-Fish-Little-Pond-Effect is a classic example of compositional effect (see also previous section 1.1 in the Introduction). In this respect, studies on the BFLPE hypothesis are inherently multilevel with individual units (e.g. students) nested within group-level units (e.g. schools). Then the focus is on whether the group-level effect (the class- or school- average ability) has a significant effect after controlling for the corresponding individual-level effect (individual ability). In summarizing the minimal conditions to test the BFLPE (see also Marsh, 2007b, Marsh and Craven, 2002, Marsh, Seaton et al., 2008), necessary components include:

- A multilevel design with a satisfactory number of groups (schools or classes involved) and a sample of an adequate size of students within the school (representative or total).
- An objective measure of achievement for each individual student that is directly comparable over different schools and an appropriate measure of academic self-concept.
- Tests of the effects of school-average achievement on academic self-concept after controlling for the effects of individual student achievement.

2.1.8 Methodological Advances Applied in and Stimulated by BFLPE Research

The BFLPE and, more generally, studies involving the analyses of compositional effects are inherently multilevel with individual-level units (e.g. students) nested within group-level units (e.g. the school). When the interest lies in measuring compositional effects, then Multilevel Modelling (MLM; see Goldstein, 2011; Bryk and Raudenbush, 1992; Snijders and Bosker, 2004) is normally incorporated. Nevertheless, in the initial studies investigating the phenomenon, this was largely ignored as they were based on single level models depending on manifest indicators (single test scores) at the level of individual students (Marsh, 1991, 1987, 1984; Marsh and Parker, 1984). In this way the measurement error at level-1 was also not taken into account in the analyses. Moreover, there was no control for sampling error in the higher-level measures. Despite the fact that these methodological flaws were not considered as limitations at the time these studies were conducted, there have subsequently been important methodological advances in the appropriate statistical analysis of the BFLPE. For instance, not accounting for the clustering of students within schools (or classes) could have led to underestimation of the standard errors leading to a higher probability of making a significant finding where none existed.

2.1.8.1 Multilevel Models and Structural Equation Models in Big-Fish-Little-Pond-Effect studies

The first BFLPE study to embrace a multilevel perspective and effectively take into account the nesting of students within the classes or the schools (see section 1.2) was conducted by Marsh and Rowe (1996). Subsequent research (Lüdtke, Köller, Marsh and Trautwein, 2005; Marsh et al., 2000) also incorporated a multilevel perspective, viewing students as nested within classes or schools. Large-scale international comparison studies (Marsh and Hau, 2003; Seaton et al., 2009) have been able to demonstrate the BFLPE using three level multilevel models with students at the lower-level, schools at the second level and countries at the third level of the models. Liem, Marsh, Martin, McInerney and Yeung (2013) have even conducted a multilevel study with four levels (student, class, track, school).

Despite the advantages associated with the use of multilevel models there were still serious problems that could be identified: One of the main limitations in their use as a tool for investigating compositional effects in BFLPE analyses was that they typically ignored potential unreliability in educational data. The use of Structural Equation Models (Marsh, 1994) allowed manifest variables to be replaced by latent variables with multiple indicators so that adjustments for measurement error at level 1 could be achieved. There were also attempts to make crude corrections for sampling error at level 2 (e.g. Marsh, 2007b). However, analyses were single level only- clustering in the data was not taken into account. A step forward in relation to this limitation was the study by Marsh and O'Mara (2010) that employed more advanced SEM models: the researchers used the complex design option in Mplus. While this approach allows the correct estimation of standard errors and controls for clustering effects, it fails to control for sampling error in the data.

2.1.8.2 The development of Multilevel Structural Equation Models stemming from BFLPE Research

Despite the wide use of multiple-indicator latent-variable models to control for measurement error (Structural Equation Models) and multilevel models accounting for hierarchical data, integrating the two frameworks into one – multilevel structural equation models – has been slow to evolve (Marsh, Lüdtke, et al., 2009).

Lüdtke, Marsh, et al. (2008) demonstrated how the Multilevel Latent Covariate Model could be used to assess compositional effects correcting for bias due to sampling error arising in the aggregation of lower-level units to form higher-level constructs. However, because this model used single scale variable at level 1, it did not account for measurement error due to the sampling of items. A series of studies followed, stimulated by evolving statistical methodology (Muthén and Muthén, 2008-2012) and limitations in BFLPE (Lüdtke, Marsh, et al., 2011; Marsh, Lüdtke, et al., 2009) extended this initial demonstration of a multilevel model that controlled for sampling error in level 2 constructs.

Marsh, Lüdtke et al. (2009) proposed a set of four models for investigating compositional effects, classified into a 2x2 taxonomy. An extensive discussion on the four models of the 2x2 taxonomy is given in the following section of my literature review (section 2.2). The models are multilevel in nature and thus account for the hierarchical structure in the data. The models allow partial or full corrections for (i) measurement error, using multiple indicators, at level 1, and (ii) measurement error at level 2 as well as for (iii) sampling error arising from using only a small number of level 1 units to form the higher-level aggregates. The Multilevel Latent Covariate Approach, which was demonstrated by Lüdtke, Marsh, et al. (2008), is the second model in this 2x2 taxonomy. The doubly-latent multilevel model, the fourth model of the taxonomy, not only allows adjustments for sampling error, but also for measurement error at level 1 and level 2.

Research using these methodological advances in the field of educational psychology and self-concept research (see for example Nagengast and Marsh, 2011) has been able to demonstrate BFLPEs accommodating both for the prevalence of nested structures in the data as well as for unreliability in the educational data due to measurement error (see section 1.3 on the definition of measurement error) and sampling error (see section 1.3 on the definition of sampling error). However, the application of these models is not only limited to big-fish-little-pond-effect; it can be used in evaluating group-level effects in a wide range of studies. For instance, Marsh, Lüdtke et al. (2012) demonstrate how the models can be used to assess classroom climate and context effects to address conceptual and methodological issues in relevant research. Importantly, in my thesis, I show how these models, in themselves or suitably modified, can be used in educational effectiveness research, in which the main outcome variable is academic achievement, for a more reliable assessment of the impact of school composition (see section 1.1) on students' educational outcomes.

The Marsh, Lüdtke et al. (2009) models have been extended to even more complex models: to doubly latent multiple-group models (Nagengast and Marsh, 2011; 2012), which combine multiple-group and doubly latent approaches. In my thesis, I present in detail the Marsh, Lüdtke et al. (2009) 2x2 taxonomy of multilevel structural equation models - a methodological basis for much of my thesis.

2.2 Addressing Measurement and Sampling Error Bias in Compositional Effects Estimates: The Four Models of the 2x2 Taxonomy

In this section of my literature review I describe the compositional analysis model – the conventional methodological approach to the two main substantive issues that I address in my thesis: School compositional effects of prior achievement and the Big-Fish-Little-Pond-Effect (BFLPE) hypothesis. I continue with a reference to the inability of the conventional multilevel modelling framework as used in the assessment of compositional effects to account for

measurement error in the level 1 and level 2 variables and for sampling error in the aggregation of the level 1 constructs to form the level 2 constructs. I explain how this limitation of multilevel modelling induces bias in compositional effects estimates. The section ends with a presentation of the four models of the Marsh, Lüdtke et al. (2009) 2x2 taxonomy; these models are used comprehensively in my thesis to estimate compositional effects correcting for measurement and sampling error bias.

2.2.1 Compositional Analysis Models

Compositional analysis is the methodological tool typically used in educational research when investigating whether school, classroom or teacher (level 2, higher-level, group-level) aggregates of student-level (level 1, lower-level, individual-level) characteristics contribute to the prediction of students' outcomes, over and above what can be explained by the individual-level variable on which aggregation is based (e.g. in school and teacher effectiveness studies, value added models, school or classroom climate and contextual studies).

Assuming that there is a two-level structure in the data, with individuals nested within-groups (e.g. students nested into schools) and that an individual-level variable X (e.g. a student's achievement) predicting a dependent variable Y (e.g. students' subsequent achievement or academic self-concept), then the two-level model, conventionally used for the investigation of compositional effects, can be expressed in the following way (the notation is adopted by Snijders and Bosker, 2004):

$$\text{Level 1: } Y_{ij} = \gamma_{oj} + \gamma_{10}(X_{ij} - \overline{X}_{.j}) + R_{ij} \quad (2.1)$$

$$\text{Level 2: } \gamma_{0j} = \gamma_{00} + \gamma_{01}\overline{X}_{.j} + U_{0j} \quad (2.2)$$

The equations given in relationship (2.2) can be combined with equation (2.1) giving:

$$Y_{ij} = \gamma_{00} + \gamma_{10}(X_{ij} - \overline{X}_{.j}) + \gamma_{01}\overline{X}_{.j} + U_{0j} + R_{ij} \quad (2.3)$$

In relationship (2.1), Y_{ij} is the outcome for person i in group j. The individual-level predictor X_{ij} is centred around the group mean $\overline{X}_{.j}$ (group-mean centring[2]). In relationship (2.2) the intercept γ_{0j} is the dependent variable. Relationship (2.3) describes a random intercept model—only the intercept is allowed to vary across groups.

Group mean centring the predictor variable implies that γ_{01}, the slope relating the random variable $\overline{X}_{.j}$ to the intercepts from level 1 equation is the between-group regression coefficient, describing the relationship between the aggregates $\overline{X}_{.j}$ and $\overline{Y}_{.j}$. The fixed effect of the individual-level variable X on Y, (γ_{10}) is the within-group coefficient describing the relationship between X_{ij} and Y_{ij} within each group (For a more detailed discussion on within- and between- regressions see Snijders and Bosker, 2004). To obtain the compositional effect—that is, the effect of the group average $\overline{X}_{.j}$ on Y_{ij}, the difference γ_{01}-γ_{10} should be used: a compositional effect is present if γ_{10} is significantly different from γ_{01}.

The parameter γ_{00} represents the overall mean of the outcome across all individuals and across all groups, while the quantity U_{0j} represents the residual at the group-level—this is

[2] An equivalent way of specifying a compositional analysis model would be to use grand mean centring of the predictor by subtracting from each level 1 observation the grand mean of the level 1 predictor. Lüdtke et al. (2008, pp. 206-207) explain how the compositional effect can be obtained by the grand mean centring approach, and the equivalence between the grand mean centring approach and the group mean centring approach.

denoted by allowing only the indicator for the group, j, to appear in the subscript of the residual, showing that the residual can only vary between groups. Its variance represents variability between groups that is not explained by the explanatory variable. The residuals at the within-level are given in the relationship (2.1) by R_{ij}. In a similar way, they indicate the extent of variability between individuals not explained by the explanatory variable. The residuals U_{0j} and R_{ij} are assumed to be independent from each other and are normally distributed, with mean zero and variance τ^2 and σ^2 respectively.

To sum up without reference to the technical details, a multilevel compositional model can be described by the following main components:

(1) the effect of the individual-level covariate, X, on the individual-level outcome, Y (the "within-group", "individual-level" or "level 1" effect).

(2) the effect of the higher-level aggregate, $\overline{X}$, on the individual-level outcome, Y (the "compositional" effect).

(3) a residual at the individual-level, that expresses the variability at level 1 in the outcome, Y, after adjustments for the individual-level covariate, X and the corresponding aggregate, $\overline{X}$. This residual is referred to as the "within-group" residual or "level-1" residual or "individual-level" residual. In an analogous way, its variance is referred to as the "within-group" (residual) variance or "level-1" (residual) variance or "individual-level" (residual) variance.

(4) A residual at the group-level, that is indicative of the variability at level 2 in the outcome, Y. This residual is referred to as the "between-group" residual or "level-2" residual or "group-level" residual. Similarly, its variance is referred to as the "between-group" (residual) variance or "level-2" (residual) variance or "group-level" (residual) variance.

The multilevel model which is typically used for the estimation of compositional effects involves the use of single manifest indicators at the individual-level while group-level variables are computed as the simple average of individual-level variables within each group. Thus there is no control for measurement error at the individual-level and group-level or for sampling error related to the aggregation of individual-level constructs to form the group-level constructs. The way this introduces bias into the estimates of compositional effects is illustrated in the next section.

2.2.2 The Prevalence of Measurement and Sampling Error Bias in Compositional Analysis Estimates

Conventional compositional models as implemented in the multilevel modelling framework have one major limitation: They fail to take into account the presence of measurement error in the underlying data. This can be particularly problematic considering that measurement error in the educational data is the rule rather than the exception; measures of students' performance are never perfectly reliable (Warner, 2008). Moreover, particularly when higher-level aggregates are involved in the analysis, another source of error can be distinguished: sampling error, the result of aggregating observations from only a small number of lower-level units to form the higher-level unit. This may also have consequences for the compositional analysis estimates.

The inferences that I make here are based on empirical studies both from the field of self-concept research (e.g. Marsh, Seaton et al., 2010) as well as from the field of educational effectiveness (e.g. Burstein, 1980; Gray et al., 1990; Harker and Tymms, 2004; Hutchinson, 2004; Hutchinson, 2007; Woodhouse at al., 1990). They have also been verified by mathematical formulas which have been derived based on estimates of reliability in the educational data separately for individual and group-level measures within the Marsh, Lüdtke et al. (2009)framework (see Marsh, Lüdtke et al., 2012). I refer the reader to Appendix A for an overview of these statistical derivations (see, specifically, section A.2).

Here I restrict the discussion to the practical implications that can be inferred based on these formulae; in my view this is what a researchers should be aware of when conducting empirical research.

2.2.3 Measurement Error and Sampling Error Bias in Compositional Analysis Estimates

2.2.3.1 Measurement error in explanatory variables at level 1

A well-known consequence of measurement error in explanatory variables in single-variable situations is the attenuation of their estimated coefficients (Hutchison, 2004). However, when there is more than one predictor involved in the analysis, failure to control for unreliability in one predictor may misleadingly increase or decrease the path coefficient of another variable, leading to a positive or a negative bias (Marsh, Seaton et al., 2010).

In compositional models (see previous section 2.2.1) the outcome variable, say Y, is typically regressed on (at least) two variables controlled for in the model: the individual-level measure, say X and the corresponding aggregate, say, $\overline{X}$. Consider for example, the compositional model used for testing the big-fish-little-pond-effect hypothesis. In this model, X is the students' prior achievement, $\overline{X}$ is the school average achievement and Y is students' self-concept. Consider also the compositional model used for the assessment of school compositional effects in educational effectiveness research (e.g. in value added analysis), that I also consider in my thesis. In this model, variable X is prior achievement, $\overline{X}$ is average prior achievement and Y is subsequent achievement. Then, the two predictors (X and $\overline{X}$) are obviously positively correlated to each other, while the individual-level variable typically has a positive effect on the outcome variable Y. In this scenario, increasing measurement error variance at level 1 will attenuate the estimated within-group effect, causing a negative bias in its estimation – one of opposite sign of the original within-group effect. At the same time, a positive bias, one of the same sign of the within-group effect will occur on the estimated effect

of the corresponding aggregate. In this way, there are three different possibilities, with measurement error in individual-level measures (X) leading to bias in the estimated compositional effects in the following ways (see also Marsh, Seaton et al., 2010): (i) true compositional effects misleadingly appearing larger or smaller. For instance, when the compositional effect is originally positive, it becomes inflated, that is more positive. On the other hand, when the compositional effect is originally negative, it appears to be smaller in absolute value, that is, closer to zero; (ii) apparent positive compositional effects, even if their true value is zero; (iii) apparently positive compositional effects, even if their actual value is negative. However, it is important to note that all three of these possibilities represent a positive bias to the estimated compositional effect.

2.2.3.2 Measurement error in explanatory variables at level 2

The prevalence of measurement error in explanatory variables at level 2 has an impact on the estimation of only the fixed effects of level 2 variables; it does not affect the estimated within-group effect. Specifically, it causes attenuation in the estimated effect of the corresponding aggregate. Whenever relevant, it can also have an impact on the effects of other level 2 variables controlled for in the models (Woodhouse et al., 1996).

2.2.3.3 The impact of measurement of measurement error in explanatory variables on standard error estimates

Measurement error in the predictor variables in compositional effects leads to apparently smaller standard errors, so that controlling for measurement error in the corresponding variable leads to larger standard error estimate (see Ferrão and Goldstein, 2008, Woodhouse et al., 1996). Still, the impact of measurement error on standard error estimates is not as severe as on the actual coefficient estimates (see also, Hutchison, 2004).

2.2.3.4 Measurement error in the outcome variable

Based on empirical results, Ferrão and Goldstein (2008) claim that the impact of allowing for measurement error in the response variable in regression models is the same as when allowing for measurement error in the predictor (increased level 1 coefficients, larger standard error estimates) but, they note that the impact is not so severe. What can, in fact, be shown (see section B.3 in Appendix B) using theoretical derivations is that measurement error in the outcome measures of single and multilevel regression models does not actually cause bias in the estimated coefficients per se; it affects only standard errors, leading to larger standard error estimates.

2.2.3.5 Measurement error bias in random effects

The interest in compositional analysis models is not only on the impact of measurement error at level 1 and level 2 on the estimated within effect and on the compositional effect of the aggregate. The implications of measurement error for the estimates of the random effects can also be of interest. Woodhouse et al. (1996) explain that adjustments for measurement error at level 1 (whether this adjustments refer to the individual-level predictor in the models or in the outcome) can cause smaller estimated level 1 residual variance but almost unchanged level 2 residual variance. Further adjustments for measurement error at level 2 may lead to even smaller level 2 residual estimates. However, this change is small compared to the reduction in the level 1 variance. It should be noted that the effects of measurement error on the residual terms of compositional models are more prevalent when no adjustments are made for the school-level aggregate. When compositional effects are included in the models, they can compensate for level 1 measurement error (off-setting biases) so that the residuals are not much affected although the level 1 and level 2 estimated effects may be biased by measurement error.

2.2.3.6 Bias due to sampling error

The direction of bias in compositional effects estimates due to sampling error depends, among other factors, on whether the actual effect is positive or negative. On the basis of mathematical derivations, Lüdtke et al. (2008) suggest that when adjustments are not made for sampling error in compositional analysis that assumes latent aggregation (as in the present study; see section 2.2.5), negative compositional effects are estimated to be less negative, closer to zero, while positive compositional effects are underestimated (estimated to be less positive, closer to zero). In other words, the relation between the group-level construct and the individual outcome is shown to be weaker when sampling error is prevalent in the data. This bias is larger for a smaller group size and for less agreement in the observations within the same group (lower intra-class correlation, *ICC* ; see Appendix A)

2.2.4 Compositional models: A Methodological Tool with Important Substantive Applications in Self-Concept and Educational Effectiveness Research

The present section has focused on the methodological implications of inadequate controlling for measurement and sampling error in compositional analysis. However, this inadequacy in the modelling of compositional effects can also have substantive implications: It can result in invalid substantive inferences when these are based on estimates obtained with multilevel compositional models. An example of an issue of high substantive relevance for educational psychology and self-concept research, with its methodological underpinnings lying on compositional analysis, is the Big-Fish-Little-Pond-Effect hypothesis. In section 2.1 I have explained what the Big-Fish-Little-Pond-Effect paradigm entails and how recent methodological developments in this area of research have been able to address measurement and sampling error bias in compositional analysis estimates. These methodological advances can be applied in educational effectiveness to address on-going debates on school compositional effects on students' outcomes. I refer to this issue in another section of my literature review

(section 2.3): "School Compositional Effects in Educational Effectiveness Research: Using the Marsh, Lüdtke et al. (2009) 2x2 Taxonomy to Solve Long-Standing Debates". In the following section I describe the Marsh, Lüdtke et al. (2009) framework.

2.2.5 The Marsh, Lüdtke et al. (2009) 2x2 Taxonomy of Multilevel Structural Equation Models

The Marsh, Lüdtke et al. (2009) framework integrates the two foremost approaches in educational research, multilevel modelling and structural equation modelling. Marsh, Lüdtke et al. proposed a set of four multilevel structural equation models which they classified into a 2x2 taxonomy (see Table 2.1) based on the corrections made for measurement error and for sampling error.

All of the four models can be characterised as compositional models, taking the multilevel nature of the data into account, with the level 2 unit being the school (or the classroom) and the level 1 unit being the student. In their structural part, each of these models can be described by a regression equation, in which the criterion (e.g. self-concept in big-fish-little-pond-effect studies) is regressed on an individual-level variable (e.g. prior achievement) and the corresponding aggregate (e.g. school average achievement) – at the very least. They all have a random component at the higher-level (e.g. the school) and at the lower-level (e.g. the student). What makes these models distinct from each other is the way they conceptualize the individual-level and group-level variables that are used in the modelling. Some of them view level 1 variables as manifest, measured by a single scale score and, therefore do not make adjustments for measurement error. Some view them as latent, measured by multiple indicators, and in this way allow for multivariate corrections for measurement error. With respect to the group-level constructs, a subset of the four models assumes manifest aggregation, not controlling for sampling error, while the remaining models assume latent aggregation, allowing for sampling error adjustments.

In this section, I present the unique way of approaching compositional analyses by the four models of Marsh, Lüdtke et al. (2009) 2x2 taxonomy. For a more technical presentation of the four models I refer the reader to Appendix A (see, specifically section A.3). A summary of the models is also shown in Table 2.1; the reader should refer to this table to see the adjustments made by each of the four compositional models.

2.2.5.1 The "Doubly Manifest" approach: The conventional Multilevel Modelling approach

The first model in the 2x2 taxonomy (doubly manifest) is the conventional multilevel modelling approach (see 2.2.1) which involves the use of single manifest indicators as measures of the individual-level variables – both the predictor and the outcome variable. When item level data are available at the level of the student, then the student-level indicators are obtained by taking the average score across the different items - the school-level aggregate is formed as the mean score of students within the school. This approach is manifest in relation to the sampling of items (no adjustments made for measurement error at either level 1 or level 2) and manifest in relation to the sampling of people (no adjustments made for sampling error).

2.2.5.2 The "Manifest Latent" approach: making adjustments for sampling error assuming latent aggregation

The second approach (manifest latent) uses scale scores for the student-level variables. It is therefore also manifest in relation to measurement error. It differs from the first model in that it uses latent rather than manifest aggregation for the construction of the school-level aggregate and, in this way it takes sampling error into account.

2.2.5.3 Distinguishing between a formative and a latent aggregation process

To understand where the difference lies between the doubly latent approach and the manifest latent approach, a distinction should be made between two different sampling processes underlying the construction of an aggregated construct: Formative (manifest) aggregation and reflective (latent) aggregation.

A formative (finite or manifest) aggregation process is based on aggregating observations from a group with an underlying finite population, assumed to consist of the total number of units within the group. The resulting aggregate can then be characterized as a formative construct. Sampling error is prevalent in the construction of formative constructs when the aggregation involves only a sub-sample of the units within the cluster. This type of sampling error was addressed, for example, by Woodhouse et al. (1996) who showed how to make adjustments for measurement error and for sampling error when the observations represent a sample within each cluster and their means are just a sample estimate of the true cluster means.

Reflective (infinite or latent) aggregation is assumed to be based on an infinite sampling process. Group characteristics are latent, unobserved constructs that can be inferred on the basis of a finite subset of a potentially infinite number of observers (Marsh et al., 2012): The value of the aggregated construct is assumed to be just a manifestation of all the potential values that it could have taken based on finite samples obtained in different ways by the infinite population. For reflective constructs then, sampling error is prevalent in the aggregation of individual-level units to form the group-level construct, even if all the units within the cluster are used, as these are simply a sample of a larger, potentially infinite population of individuals. This is the type of sampling error discussed by Lüdtke et al., (2008) and assumed by the second model of the 2x2 taxonomy to adjust for sampling error in the school-level aggregate.

2.2.5.4 The "Latent Manifest" approach: controlling for measurement error through the use of multiple indicators

The third (latent manifest approach), adjusts for measurement error, since multiple indicators are incorporated to measure attainment. However, manifest aggregation is followed to form multiple level 2 indicators from these level 1 indicators: The school-level indicators are formed by just taking the observed school-level average of the corresponding individual-level variables. Thus, no adjustments for sampling error are made.

2.2.5.5 The "Doubly Latent" approach: A full correction approach

The fourth approach (the doubly latent approach) makes adjustments for both measurement error and sampling error. Multiple indicators are used to infer the value of the student-level outcome as well as the student-level predictor in the models. The school-level variable is also based on multiple indicators and reflective aggregation is followed: Instead of taking the observed school average score as in the latent manifest approach, each indicator is decomposed into a latent within and between parts in a multilevel Confirmatory Factor Analysis (CFA) framework. The latent between group components of the observed indicator are then used as the level 2 indicators of the group-level construct.

There is an analogy in the way the doubly latent model uses multiple items to control for measurement error in averaging items to obtain the students' scale scores and the way in which it incorporates multiple level 1 scores from students within the same cluster, to control for sampling error in the aggregation of level 1 scores to from the latent level 2 variables (see also Marsh, Lüdtke et al., 2012): In doubly (and partial) latent models, estimates of measurement error are smaller when the correlations among multiple indicators are higher, suggesting that agreement among the items is better and when there is a larger number of items - the traditional approach to reliability analysis. Moreover, a minimal requirement for support of the latent factor structure is that agreement among the different items designed to measure the

same factor is higher among items corresponding to different factors. In a similar way estimates of sampling error are smaller in the prevalence of higher agreement among different individuals in the same groups and when the number of individuals within each group is larger – just like the traditional measures using intra-class correlations (see Snijders and Bosker, 2004). A minimal requirement for support of the latent higher-level construct is that agreement among the students within the same group is higher than agreement among students from different groups.

The partial (manifest latent and latent manifest) and full correction (doubly latent) approaches discussed in this section should be applied with suitable caution. The doubly latent model is not a panacea; on the contrary, under certain circumstances the partial correction approach or even the doubly manifest approach should be used for the investigation of compositional effects (see sections 6.2 and 6.3 in the discussion where I explain which approach should best be used).

Table 2.1: A 2x2 taxonomy of Compositional Models[1]

<table>
<tr><td colspan="4">Set of Four Core Compositional Models</td></tr>
<tr><td colspan="2"></td><td colspan="2">Sampling of Persons

(Sampling Error)</td></tr>
<tr><td colspan="2"></td><td>No</td><td>Yes</td></tr>
<tr><td rowspan="2">Sampling of Items (Measurement Error)</td><td>No</td><td>Doubly Manifest

Approach:

• Single manifest indicators (one per factor)
• Manifest aggregation of L1[2] constructs to form L2 constructs</td><td>Manifest Latent

Approach:

• Single manifest indicators (one score per factor)
• Latent aggregation of L1 constructs to form L2 constructs</td></tr>
<tr><td>Yes</td><td>Latent-Manifest

Approach:

• Multiple Indicators (constructs are latent in relation to items)
• Manifest aggregation of L1 indicators to form L2 indicators</td><td>Doubly Latent

Approach:

• Multiple Indicators (constructs are latent in relation to items)
• Latent aggregation of multiple L1 indictors to form multiple L2 indicators</td></tr>
</table>

Note.[1] Adapted from "Doubly-Latent Models of School Contextual Effects: Integrating Multilevel and Structural Equation Approaches to Control Measurement and Sampling Error" by. Marsh, Lüdtke et al., 2009. [2]In this Table, L1 is used to denote student-level (level 1) variables and L2 is used to denote school-level (level 2) variables.

2.3 School Compositional Effects in Educational Effectiveness Research: Using the Marsh, Lüdtke et al. (2009) 2x2 Taxonomy to Solve Long-Standing Debates

In my thesis (see Study1) I consider school average achievement, which is a widely used school-level (level 2) construct in the educational literature. I investigate the impact of school average mathematics achievement on students' subsequent mathematics achievement over and above the effect of individual-level prior mathematics achievement. In other words I seek to answer the question: "Does the ability composition of the school in which a student finds himself/herself in terms of average level of academic achievements exert an influence on students' outcomes (mathematics achievement and self-concept) over and above what can be explained by the own individual's characteristics as measured by his/her prior achievement?"

In this section I review the literature, both theoretical and empirical, in relation to the incidence of school compositional effects in educational settings. I begin with highlighting the relevance of compositional models as a methodological tool with substantive issues within the educational effectiveness paradigm: compositional analysis is the conventional statistical approach used in assessment of the effect of school composition (Nash, 2003). I define "school compositional effects", explaining how they can be interpreted practically. A number of reasons underlie the occurrence of school compositional effects in educational settings; these are also discussed in the present section and the connection of these issues to value added models used in educational effectiveness research for school accountability (see section 2.3.1 for a definition of educational effectiveness research) is highlighted. I then review existing literature on the wide disagreement in relation to the nature and the size of school compositional effects on students' outcomes. In part, this on-going debate is due to statistical problems associated with the measurement of individual-level variables (see, for example, Harker and Tymms, 2004) or of the corresponding school-level aggregates (see Lüdtke et al., 2008). Such inadequacies in the statistical procedures can lead to misleading, or at least biased, compositional effects.

Nevertheless, it is crucial to be able to measure the school composition as accurately as possible in order to be able to differentiate between the effects of the school composition from the effects of school processes on student achievement.

2.3.1 The Use of Compositional Analysis for the Assessment of the Effect of School Composition: A Substantive Application in School Effectiveness Research

"School Effectiveness Research" (SER) is concerned with the impact of school-wide factors (e.g. school policies on teaching or the school composition) on students' cognitive and affective performance (Creemers, Kyriakides and Sammons, 2010). It seeks to distinguish between the effect of the background that students bring to school and the effect of their educational experiences at the institution, whilst recognizing that there are a number of factors other than the school which can affect students' outcomes (Sammons and Bakkum, 2011): individual characteristics (such as age, gender), family socio-economic characteristics (for example, the socio-economic status of the student's family) and community and societal characteristics (e.g. the neighbourhood context). The term "Educational Effectiveness Research (EER)" has recently emerged (see Creemers and Kyriakides, 2008; Muijs, 2006; Teddlie, 2010) superseding the more traditional term "School Effectiveness Research" (Muijs, Kelly, Sammons, Reynolds and Chapman, 2011). This is to reflect the more general mode of enquiry in the field, addressing, for example, teacher effectiveness as well as school effectiveness.

In the classic studies that led the development of the school effectiveness paradigm (e.g. Coleman, Campbell, Hobson, McPartland, Mood, Weinfeld and Robert, 1966; Jencks, Smith, Ackland, Bane, Cohen, Ginter, Heyns and Michelson, 1972), there appeared to be modest differences in the student outcomes that could have been attributable to their educational environment.

Although school effectiveness research grew in a climate hesitant to accept that schools can make a difference to children's development, there is now a widespread assumption internationally that schools do affect children's learning; Indeed, they "add value" (Teddlie and Reynolds, 2000). As stated by Reynolds and Creemers (1990, p.1), "… simply, schools do make a difference".

Close to the notion that "School Matters" (Mortimore, Sammons, Stoll, Lewis and Eccob, 1988) is the interest in assessing the quality of the schools. For this reason, different measures of school effectiveness have been developed in the literature reflecting both relative differences in their effectiveness across schools (relative school effects) as well as absolute schooling effects - the question for the latter being "Does schooling have an effect"? (Wiley and Hamischfeger, 1974, p.7).

The focus of enquiry then becomes, what is the actual nature of school effectiveness? That is, why do schools appear to have their own different effects? According to Thrupp and Hursh (2006; see also Lauder, Kounali, Robinson and Goldstein, 2010) two apparently opposing viewpoints can be identified in the literature: The first perspective is that school effectiveness varies only as a function of the schools' management and the quality of the teaching. The second standpoint is that pupil composition of the school and other social factors also determine the pupils' outcomes.

The extent to which the factors relevant to the school composition affect student outcomes is typically investigated using compositional analysis (see section 2.2.1): If a student's performance in a school is indeed affected by the characteristics of his or her fellow students (Marsh, Kong and Hau, 2000; see also Lüdtke et al., 2008), this gives rise to the predominance of school compositional effects.

2.3.2 School Compositional Effects: A Definition

Harker (2004) gives a good definition of the school compositional effect:

> "A compositional effect is said to exist when a variable (such as the students' socio-economic status) as an aggregated variable at the school-level makes a significant contribution to the explanatory model over and above the contribution of the same variable in the model at an individual-level." (Harker, 2004, p. 2).

In my thesis (see Study1) my focus is on compositional effects of school average achievement. I investigate the impact of school average mathematics achievement on students' subsequent mathematics achievement over and above the effect of individual-level prior mathematics achievement.

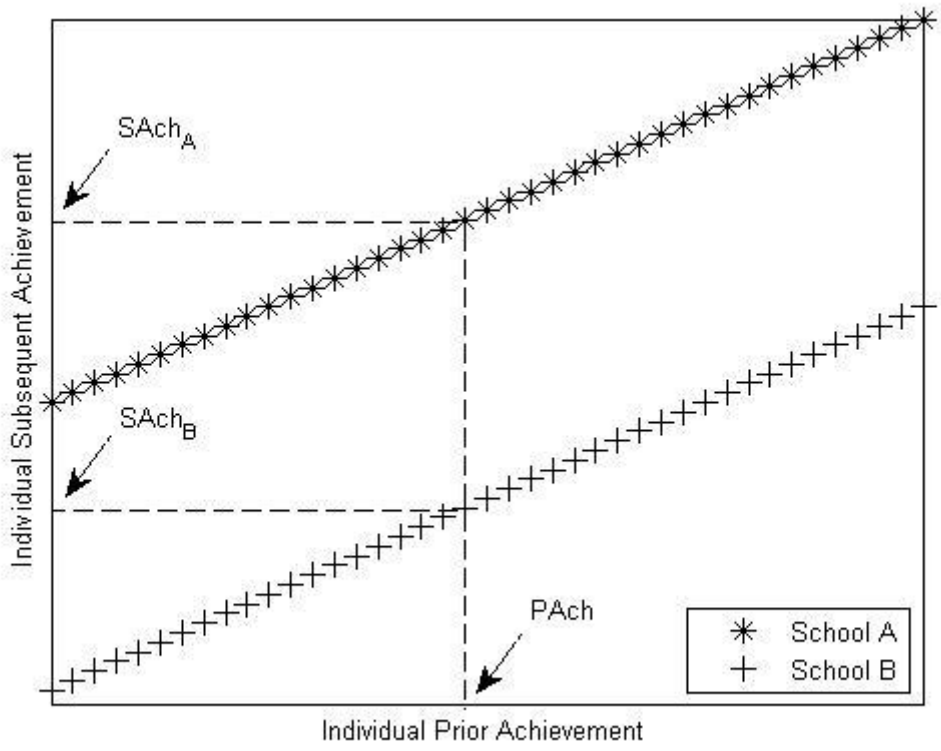
2.3.3 Interpreting School Compositional Effects

Whenever a school compositional effect of prior achievement is identified in the data, it means the extent to which a school raises a student's performance depends on average achievement of the student's school peers. In this case, a student is expected to achieve higher (or a lower) over and beyond what is predicted by his (or her) individual prior achievement; the difference between the predicted and the actual achievement of the student depends on the average achievement of the students in the school that the student attends.

Suppose now that a positive school compositional effect of school average achievement is detected. If two schools, say school A and school B, are randomly selected from my sample then the linear lines representing the relationship between prior achievements versus subsequent achievement for the students within each school separately would look like those depicted in Figure 2.3. School A (see Figure 2.3) is the one with higher average achievement while school B is that with lower average achievement. Thus, a student with a given prior achievement is expected to have a higher subsequent achievement if he or she attends school A as opposed to what he/she would achieve when attending school B.

To understand this better, imagine drawing a vertical line from any given point ($PAch$) on the horizontal axes in the graph depicted in Figure 2.3– this point represents the prior achievement of a given student. The point at which this line cuts the regression line for school A ($SAch_A$) corresponds, on the vertical axes, to the expected subsequent achievement of the student given that he or she attends school A. On the other hand, the point at which this line cuts the regression line for school B ($SAch_B$). corresponds to the expected subsequent achievement of the student given that he or she attends school B. Obviously, students that attend school A are expected to achieve systematically higher than those attending school B given that they have the same individual prior achievement. If a negative school compositional effect of average achievement is detected, then the regression line describing the relationship between prior and subsequent achievement for school A would be parallel with the line for school B but would appear lower in the diagram.

Figure 2.3: An illustration of how to interpret a significant positive or negative school compositional effect



Note. In Figure 2.3 $PAch$ is a given students' prior achievement, $SAch_A$ is the students' subsequent achievement if the student attends School A while $SAch_B$ is the students' subsequent achievement if the student attends School B. The assumption is that the school average achievement in School A is higher than the school average achievement in School B while a positive compositional effect of school average achievement is prevalent in the data.

Note that in giving the interpretation of school compositional effects I have assumed that the linear relationship between prior achievement and subsequent achievement within each school is the same (this is also in accordance with the compositional model that I have given in section 2.2.1).

2.3.4 Reasons Underlying the Occurrence of School Compositional Effects

The investigation of school compositional effects is of great significance both at political and personal level, as well as from a theoretical perspective (Willms, 1985b), not only for education but for other disciplines as well (e.g. educational and organizational psychology, health, sociology and econometrics; for example, see Bliese, 2000). Especially in EER, the study of compositional effects helps in responding to parents' questions such as "What difference would it make to my children's education if they attended a school with a different composition from the one that I am thinking that he/she will go to?" or questions posed by policy–makers, for instance: "What difference will it make to the educational system as a whole if we introduce policies designed to alter the mixing of the schools?" (Harker and Tymms, 2004).

It should be noted, nonetheless, that compositional effects should be interpreted very carefully. The theoretical stance often taken (Wilkinson, Parr, Fung, Hattie, and Townsend, 2002) is that the way in which students are assigned to classes or schools has only indirect effects on student learning (Barr and Dreeben, 1991; Hattie, 2002). Specifically, the appearance of compositional effects of achievement at the level of the school, the focus of the present investigation, is not always due to the way in which pupils of different achievement levels make up the composition of the schools. The "peer" networks that are often assumed to be the cause of compositional effects (e.g., Harker and Tymms, 2004; Willms, 1985b) may often be confounded with the effects of school organizational processes and teaching practices. It is true that in schools with an academically selective intake individual students have the opportunity to

interact with higher attaining peers and those with higher socioeconomic status: establishing friendships with them, sharing hobbies, books, and out-of-class activities. Naturally, this can have a substantial impact on their educational outcomes. Nevertheless, school compositional effects may also derive from differential teacher practices (see Campbell, Kyriakides, Muijs, and Robinson, 2003) or from differences in school organization and management processes (Scheerens and Bosker, 1997). The latter relate to the extent to which the school deals with disciplinary problems, to the amount of time that the school devotes to planning and monitoring and to the efficiency of the school's daily routines. Teacher practices refer to the extent to which the schools have teaching resources, able teachers and to the level of the texts.

The suggestion that school compositional effects may reflect differential teacher practices or diverse school or classroom resources as well as peer effects has been expressed elsewhere (Hattie, 2002; Wilkinson, 2002; Parr and Townsend, 2002).

An example of a study that addresses the way in which compositional effects should be interpreted is one by Hattie (2002). Based on the most comprehensive set of meta-analyses ever undertaken to determine what make a difference in terms of educational effectiveness, Hattie raises concerns as to whether any evaluation of practices such as tracking or ability grouping based on the investigation of compositional effects on learning outcomes are really valid. He suggests that tracking per se has minimal effects on learning – he found an effect size of about .1. He claims that at best tracking benefits only the most advantaged students and that any positive effects are in fact largely due to teaching related instructional innovations: tracking has only minor effects on learning and these are confounded with other differences. In his study, Hattie makes the point that the composition of a class or a school affects only the probability that differential instruction and learning occur. For example, in classes where students are grouped according to their ability, teachers are more likely to alter the nature of their instruction, thus providing teaching of better quality, something that overcomes the effects of tracking. Most importantly, Hattie stresses that it is essential to separate gifted educational

programs from high ability tracks since the positive effects of gifted programs are in fact likely due to changes in the curriculum and quality of education rather than to ability tracking per se.

Conclusively, compositional effects may reflect not only the way in which students react to their peers but also how they react to school systems and teacher practices, since such processes are not independent of the student body (Harker and Tymms, 2004). Therefore, even when school compositional effects are identified, they should always be interpreted very carefully (Wilkinson et al., 2000).

2.3.5 School Compositional Effects and Value Added Models of Educational Effectiveness

Despite this disagreement as to whether or not school compositional effects reflect purely the effect of school composition, this is often the assumption when school aggregates of student-level characteristics are used in value added models of educational effectiveness.

2.3.5.1 Value Added Models of educational effectiveness and relative school effects

Value added modelling applies a class of longitudinal techniques that are intended to take into account all those determinants of a student's achievement (e.g., the student's age, socioeconomic status, ethnicity, level of parental education, and fluency in English) and to isolate the contribution that schools make to the achievements of the students attending them (Sammons, 1999). Needless to say, these techniques require adjustments for individual prior achievement – the criterion being the students' achievement at the end of the period over which the effectiveness of the institution is to be assessed.

With value added models of educational effectiveness it is possible to obtain relative school effects, that is, estimates of the size of differences between between schools in their students' achievement outcomes (achievement at the second measurement point), after adjustments for achievement at the first point (Sammons and Bakkum, 2011). The intention is to

obtain a measure of the differences between schools in the average value that they add to their student's progress, not from any other influences that may also influence student outcomes.

The current statistical approach to implement value added modelling for school assessment in England and one of the most commonly used world-wide is multilevel modelling (Bryk and Raudenbush, 1992; Goldstein, 1995; Snijders and Bosker, 2004; also see section 2.2.1; paragraph A.1 in Appendix A). With the use of multilevel value added models, school effects have become synonymous with the residuals in student achievement scores situated at the school-level and the importance of school effectiveness has been determined from the percentage of the residual variance situated at that level.

Using multilevel models, school effects have been estimated to explain five to fifteen per cent of the total variance in the students' achievement (Scheerens and Bosker, 1997, Kyriakides and Luyten, 2009; Luyten, 2006). This percentage depends on what variables are included in the models as explanatory variables (see, among other studies, Coe and Fitz-Gibbon, 1998) with the magnitude of school effects being estimated as smaller the more complex the incorporated models are (Marsh et al., 2012).

2.3.5.2 The use of school compositional effects in value added models of educational effectiveness

When school compositional effects are included in value added analysis in addition to the student-level background characteristics, then the term "compositional value added models" is used to characterize the type of models produced. An example of such a model is the "Contextual Value Added Model" that was until very recently being used in England for the construction of the league tables (the E-CVA model; see Ray, 2006). The school aggregates are then conceptualized as the effects of school-level factors that may potentially influence achievement but over which the school has no control at all. These are often differentiated from

the effects of the school practices and policies - despite the fact that, as I have already explained (see section 2.3.4), this is debatable.

If compositional effects do exist and if they are significant, then taking them into account in the value added models may drastically change the estimation of relative school effects (Dumay and Dupriez, 2008; Harker and Nash, 1996; Hutchison, 2004). In a relevant discussion, Raudenbush and Willms (1995; see also Raudenbush, 2004; Hutchinson, 1993) distinguish conceptually between what they call "Type A" and "Type B" school effects. The former are obtained after adjustments for student-level characteristics only. They are of interest to parents when it comes to choosing a school for their child. The so-called "Type B" effects are obtained after controls are made for the overall ability or the social class composition of the schools, that is, for school compositional effects. They seek to answer the question of how well a certain school performed relative to other schools with similar school intakes, compositional effects and wider social influences. Hence they seek to express the influence of schools on their students through factors under the control of the school, such as school policies and practices, the administrative leadership of the school, the curricular content, the utilization of resources and classroom instruction. This is of particular importance to administrators and politicians who want to hold school-site personnel accountable for schooling outcomes and also to policy makers who are interested in discovering the best practices that may be used by schools in order to improve the school outcomes (Raudenbush and Willms, 1995; Raudenbush, 2004).

Since the effects of the school's composition are confounded with the effects of the school's policies, in the way this is discussed in section "Reasons Underlying the Occurrence of School Composition Effects", "Type A" and "Type B" effects are inseparable in practice. Therefore, the distinction between these two types of effects is not straight-forward in that there may be an overlap between Type A and Type B effects. Differentiating between the two types of effects is therefore difficult – even if they are distinct. This distinction is not particularly relevant to my thesis in that my main focus is on how to improve the estimation and appropriate

interpretation of compositional effects rather than the policy-practice issue of how these are integrated into measures of school effectiveness. However, my thesis is central to the debate in that current practice is based on sub-optimal statistical models of compositional effects – this is where I contribute to existing knowledge.

It is important to use appropriate models which are capable of taking into account potential sources of bias in the estimation of compositional effects as these are used in value added models of educational effectiveness. Only then will the school effect estimates that are based on these models have the potential to give correct information to the target audience. Suppose, for instance, that a compositional effect is found significant in value added analysis, but this is just the result of measurement error in prior achievement controlled for in these models (see section 2.2.3 and, particularly, the discussion in 2.2.3.1 on how such misleading compositional effects may occur in statistical analyses). Then the relevant statistical analysis will provide invalid confirmation of the schools' effectiveness to both parents and policy-makers. In my thesis I propose a way to correct for measurement and sampling error bias in school compositional effects whenever these are used in value added models. Although there are still many other ways in which value added models can be improved (see Appendix C in which I expand upon the limitations of value added models of educational effectiveness), this is one step towards a more appropriate modelling of the relative school effects.

2.3.6 Evidence from Previous Research on the Magnitude and Direction of the School Compositional Effect

The educational community has not yet reached a consensus on how large the compositional effect actually is; this seems to be an "enduring problem" within educational research (Thrupp, 1999; Dumay and Dupriez, 2010). Whether or not the effect is found significant or not, whether or not it is found positive or negative, seems to depend on a number of factors: the construct used to operationalise school composition, the outcome upon which the school compositional effect is considered, the statistical models employed, the nature of the sample incorporated. In

this section I review relevant studies that are themselves indicative of the wide disagreement on this topic.

I begin my review with Husen (1970) who took it for granted that students in schools with academically selective intake achieve better than those in non-selective schools. He locates the mechanism of the incidence of school compositional effects in the pedagogic environment of the school: students in selective schools find themselves in a more homogeneous environment with regard to scholastic ability which is easier for teachers to deal with and so achieve better results. Empirical evidence that supports this view can be found in the work of Willms (1985a), Barr and Dreeben (1983), Hutchinson, (1993; 2004), Teddlie, Stringfield and Reynolds (1999) and Will (1992). These studies generally suggest that students attending higher achievement institutions have an advantage over those attending schools with lower achievement intake. Even in the past Wilson (1959) demonstrated that children in schools with a higher percentage of high socioeconomic status students were more likely to attend a college than would otherwise be expected based on their status origins and academic performance (see also Alexander, Fennessey, McDill and D'Amico, 1979). Note that the construct that Wilson used to operationalize school composition related to the students' socio-economic status. However, positive compositional effects have even been reported using measures of achievement to quantify the effect of school composition. For instance, Rutter, Maughan, Mortimore, Ouston and Smith (1979) reported that pupils of similar prior attainments did better in public examinations when attending secondary schools with a higher proportion of more able pupils.

It seems that the magnitude of the school composition effect varies with the age of the students involved in the analysis. For instance, investigating the effect of the socio-economic composition on students' reading achievement, Bondi (1991) found positive school compositional effects for secondary schools but not for primary schools. Hutchinson (1993), moreover, found apparent compositional effects in older primary age pupils but not for younger pupils.

Generally, the size of school compositional effects on achievement after controlling for individual-level variables, even though positive and significant, is small. Relevant interpretations should nevertheless be made with caution: School effects of any kind are generally very small (see, for instance, Marsh, Nagengast, Fletcher and Televantou, 2010) so that although school compositional effects may be small in an absolute sense, they are not small when compared to the effects associated with other school-level variables. In this respect, school compositional effects may often be characterised as relatively large. Still, there exist studies that have found little or no evidence of school compositional effects on achievement. What is even more interesting is that some of these studies are also concerned with English achievement data (Bondi, 1991; Gray et al., 1990; Mortimore et al., 1988; Strand, 1997; Thomas and Mortimore, 1996) and are, in this way, especially relevant to the present investigation (see section 3.3.1, in which the data sample of Study 1a is described).

Importantly, although the prevalence of positive compositional effects seems intuitively reasonable, there is conflicting evidence from studies which found negative compositional effects on students' outcomes (see, for instance, Woodhouse et al., 1996). In fact, Bourdieu (1993), an influential sociologist in his lifetime, makes the point that the schools with a less selective intake are actually the ones that are more effective. Moreover, the effects of school composition are consistently negative with affective outcomes (e.g. self-concept; attitudes towards schooling), when school composition is operationalized using school average achievement (e.g. Marsh, 1987; Tymms, 2001). This relates to the big-fish-little-pond-effect hypothesis to which I have referred earlier in my literature review (see section 2.1.3).

2.3.7 A Study on the Potential Effects of School Composition: A Contribution to an on-going Debate

The wide disagreement on the effect of school composition is perhaps due to the fact that research into school compositional effects has been shaped by political, ideological and methodological concerns of the time (for instance, Harker and Tymms, 2004; Hutchinson, 2004; Lauder, Kounali, Robinson, Goldsteina and Thrupp, 2007; Thrupp, 1999; Wilkinson et al., 2000); it has never been neutral.

It was the publication of the seminal Coleman et al. (1966) report ("Equality of Educational Opportunity"), at the time of the very emergence of the school effectiveness paradigm that drew early attention to the potential effects of the school composition on the student achievement. The report was influenced by the political considerations at the time it was written when the issue of equality of opportunity was a topic of concern in America. Coleman hoped that he would find evidence of substantial inequalities in the allocation of material resources to schools of different ethnic communities. However, his findings suggested that there was little of this type of inequality. He also found that most of the school-level variables that he considered made little or no difference to school outcomes over and above the influence of student background characteristics. Nevertheless, he did report that students from ethnic minorities performed higher in ethnically integrated schools. In this way, despite the fact that Coleman's report aimed to explore ethnic inequalities, he claimed that it was the social class/prior achievement composition of the school that made the difference. Indeed, the only school-level variable that he found to have a significant impact on student outcomes was the school composition in terms of school average achievement. Coleman's conclusion that minority achievement was higher in ethnically integrated schools seemed to be inconsistent with another finding of his report: that minority students had lower academic self-concept in high-socioeconomic status schools. This last finding was in later years interpreted as the result of students suffering comparatively by having to compete with other students likely to attend

university, counterbalancing the positive normative effect on status aspirations. It supported the "frog-pond" effect hypothesis posed by Davis (1966), namely that it is better to be a big frog in a small pond than a small frog in a big pond. This hypothesis was the precursor of the "Big-Fish-Little-Pond-Effect-Hypothesis" (Marsh and Parker, 1984; see also previous section 2.1.3 in which this phenomenon is reviewed) that is also of substantial interest to the present thesis.

The interest in the effects of the school composition that was prevalent in the 1960s declined in the 1970s. A widespread belief in this period was that schools had little effect of any kind on life chances. One of the most representative studies in this period is that by Jencks et al. (1972) which claimed that not only the school policies themselves but even the composition of the school had no impact on students' outcomes. Jencks argued that earlier studies which found positive school effects of school composition suffered methodologically, in that better data and more sophisticated use of statistics would actually lead to the opposite conclusion, that is, to no virtual effects of school composition. By the late 1970s the view that schools were unable to address inequalities due to the students' background was widely accepted.

Next, during the 1980s research seems to consistently agree with the view that the success or not of a school was attributed to its climate, practices and ethos; the school policies and practices were considered to act independently of the school composition. Research in this period assumes that there exist well-performing schools that achieve academic success irrespective of their students' background and that there are certain characteristics associated with these schools to be identified by the researchers. The pupil body characteristics have been regarded as background factors, out of/beyond the school's control, that must be adjusted for in the models; the main focus lies on the factors under the school's control and how these can affect outcomes. It is also interesting that research during this period focused almost exclusively on academic achievement as the main outcome variable.

Despite the shift of focus from the effects of the school composition to the processes within the school that could make a difference to the students' achievement, the wide

disagreement in relation the nature and the size of the effect of school composition has still been prevalent (see following, see also section 2.3.6). Importantly, in recent years, an apparent interest in the potential effect of school composition has again been expressed, with studies revisiting this issue and pointing to the limitations of research addressing this issue. With the present thesis (see Study 1), I contribute to the ongoing debates by addressing limitations in the modelling of school compositional effects, and, more specifically, by proposing ways around the issue of bias due to measurement error in level 1 indicators (see section 1.7 in the introduction chapter; see also section 2.3.8 in the literature review chapter on the two types of under-specification at level 1). Moreover, I demonstrate that even when positive compositional effects are detected, they should always be interpreted with caution: they could be simply an artefact of the statistical procedures used to estimate the compositional effect (Harker and Tymms, 2004; Tymms, 1999).

2.3.8 The Prevalence of Phantom Compositional Effects: Two facets of Under-representation Biases in Compositional Effects Estimates

One of the main methodological concerns underlying the debate on the scale and the nature of the effects of school composition is the inadequacy in the design and the statistical procedures followed to detect such effects. Such inadequacy can be the result of insufficiency of level 1 covariates controlled for in the model or the existence of measurement error in the student-level variable on which the aggregate is based (see Televantou, Marsh, Kyriakides, Nagengast, Fletcher and Malmberg, in press). These are two quite distinct issues that require a separate approach for the researchers to deal with. Each one, operating through a different mechanism, has been shown to lead to bias in the estimation of compositional effects, to misleading compositional effects that are often referred to in the literature as "phantom" compositional effects (Harker and Tymms, 2004).

The way in which measurement error in the individual-level measures leads to biased estimates of the effects of the corresponding aggregates in compositional models has been described in detail in section "Measurement error in explanatory variables at level 1" of my literature review. The omission of a student-level predictor that is related both to the outcome and the school-level aggregate leads to bias in the compositional effects estimates in an analogous way. Whenever such a predictor is not controlled for in the model, then the higher-level aggregate takes over, explaining that part of the variability in the outcome which is in fact, attributed to the neglected student-level covariate.

A major focus of the present thesis (see Study 1) is the bias that arises in the estimates of compositional effects due to measurement error at level 1. Unreliability at level 1 in compositional models has been shown to lead to a systematic positive bias in the effect of the corresponding aggregate (see Marsh, Seaton, Kuyper et al., 2008) so that positive compositional effects are overestimated and negative effects are under-estimated (they misleadingly appear to be less important, close to zero). In this way, the prevalence of measurement error in individual-level measures can justifiably be claimed to be one of the main causes of disagreement in the findings regarding the size of school compositional effects.

The phenomenon of phantom compositional effects occurring due to measurement error at level 1 (the issue of measurement error bias) has been considered by numerous studies, both in itself, or in combination with under-specification due to the lack of comprehensiveness of the set of student-level variables that are controlled for in the evaluating compositional effects (the issue of omitted-variable bias): Thomas and Mortimore (1996) claimed that whenever a wide range of student-level variables are taken into account in compositional effects models, school compositional effects are no more significant in predicting student outcomes. Nash (2003) also stressed the importance of including all relevant student intake characteristics in the models and supported his view with empirical evidence using UK PISA data from year 2000. These findings echo those of Husen (1970) who implied that many apparent compositional effects

arise due to inadequate allowance for pre-existing differences. Burstein (1980) showed that spurious compositional effects arise when important student-level variables are omitted from the analysis but also raised the possibility that, even when all the correct variables are included in the analysis, misleading compositional effects arise when these variables are measured with error. Indeed, Gray et al. (1990), using data from a number of educational authorities in England, found that while some found substantial compositional effects on students' progress, others did not and the latter were ones with good measures of prior attainment. It is interesting that in his study Gray claims that most evidence in favour of the prevalence of compositional effects is, in fact, attributed to omitting key variables or using poor or inadequate measures.

Perhaps the study most relevant to my thesis concerning the two faces of under-representation biases in compositional effects estimates is that carried out by Harker and Tymms (2004; see Study 1a). In two distinct exemplary analyses, the researchers used real and simulated data to demonstrate how spurious compositional effects may arise due to issues relevant to the specification of the model or to the reliability of the predictors involved in the analyses. Using data from a New Zealand Study of secondary schools, they showed that underspecified models – in their case models that do not make adjustments for individual prior attainment - are likely to exaggerate any compositional effect. They explain that detailed and careful work is required before researchers can be fairly sure that their models are suitably structured. In an analysis that is of special relevance to the present thesis (see Study 1 and also section 1.7 in the introduction), Harker and Tymms demonstrated how an unreliable predictor can generate phantom compositional effects.

In my thesis, and, specifically, in Study 1, I replicate the Harker and Tymms (2004) study in showing how bias is induced in the estimates of compositional effects due to measurement error at level 1 (see section 3.3.5 in Study 1 where I explain how this is achieved). Importantly, while none of the studies reviewed here propose any ways around this bias, I demonstrate how this can be achieved using the Marsh, Lütke et al. (2009) multilevel structural

equation modelling framework. Using this framework (see section 2.2.5), it is possible to control for measurement error at level 1, but also for measurement error in the higher-level measures (aggregate variables) as well as for sampling error that arises when only a small number of level 1 variables are used to form the aggregate measures.

An alternative option to the Marsh, Lütke et al. (2009) compositional analyses models for evaluating the effect of school composition on students' outcomes is the use of the Regression Discontinuity (RD) approach (see Study 3 of my thesis). When the necessary hypotheses are met, the RD approach may give unbiased estimates of the absolute schooling effect – the impact of one extra year of received education through school attendance on student outcomes and the differences across schools in their absolute effects. Then it is possible to investigate the extent to which these differences across schools in their absolute effects are explained by their differences in student composition.

In the next chapter I illustrate the RD approach and the strengths that this methodological framework has over conventional compositional models. Indeed, the RD approach can be a potentially valuable tool for assessing the effect of school composition: once the main assumptions of the approach are fulfilled, there is no requirement for controls for student background variables other than age. Moreover, measurement error in student-level achievement does not bias the RD estimates - individual achievement is the criterion in RD models. Hence, the two faces of under-representation at level 1 that relate to traditional compositional analyses models are no longer relevant to RD designs.

2.4

2.5 The Regression Discontinuity Approach

To look at the quality of an educational system as a whole, i.e. to see whether formal education and schooling have an impact on students' cognitive development (see Cahan and Cohen, 1989; Wiley and Hamischfeger, 1974), educational researchers need to look at absolute measures of effectiveness. In the sections that follow, I explain how the RD approach can be used to address the need to assess the effect of schooling and education on students' achievement. Subsequently, I outline the assumptions required in order for the RD estimates to be unbiased and I highlight the advantages of the approach over conventional value added models. Lastly, I review studies that have applied the approach within the school effectiveness paradigm and that have also considered the extent to which the effect of one extra year of schooling correlated with the effect of school composition – an issue that I also address in my thesis.

2.5.1 The Regression Discontinuity (RD) Approach

The Regression Discontinuity (RD) approach is a quasi-experimental design with the defining characteristic that the probability of receiving a treatment changes discontinuously as a function of one or more underlying variables (Hahn, Todd and Van der Klaauw, 2008). People with scores below a cut-off point on a continuous variable are assigned to one treatment and people with scores above the cut-off point are assigned to another treatment. When the required assumptions are met – strict adherence to the cut-off point – unbiased estimates of the treatment effect can be obtained through appropriate modelling (Shadish, Cook and Campell, 2002, p.207-245; also see Cliffordson and Gustafsson, 2011). Other terms that have been used in the literature to characterise the RD approach are: "Regression Discontinuity Design (RDD)", "cutting-point" design (Rossi, Freeman and Lipsey, 2004, p. 289) or "between-grade level" approach (Cahan and Davis, 1987).

An important application of the RD approach in EER is its use in assessing the impact of an extra year of schooling on students' academic accomplishments. The treatment in this case is one year of schooling while the cut-off point is defined by the age at which students begin

their primary education. In this way, the RD approach is only applicable to assess the effect of one extra year of schooling only in educational systems in which students begin primary school strictly based on their date of birth. When this is the case, the RD approach can be used as an alternative to value added models to assess the schools' and teachers' effectiveness. It provides estimates of added-year effects, that is of the effect that one extra year of schooling has on students' academic attainment.

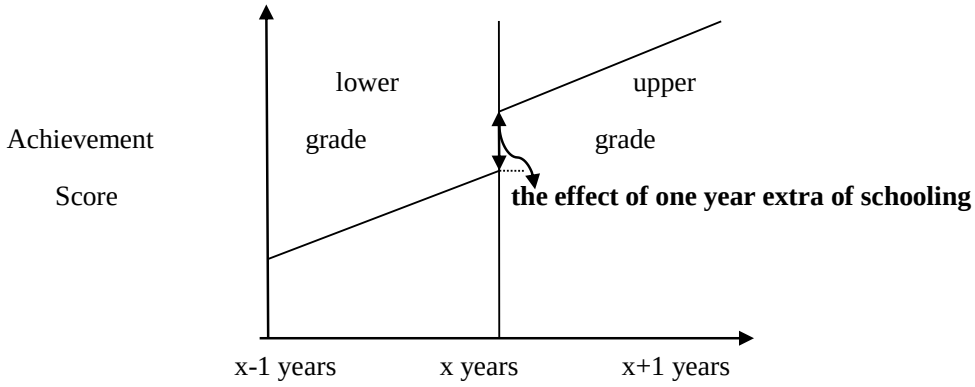
2.5.2 The RD approach as a Tool to Assess Absolute Schooling Effects

The effect size of the difference in the average attainment between pupils in two adjacent grades has been reported to be equal to an effect size (d) of .4-.5 SD with students in the higher grade outperforming those in the lower grade (Cliffordson and Gustafsson, 2011). These differences in the learning outcomes between grades could be attributed to the differences in the education received – students in the higher grade received an extra year of schooling. Nevertheless, other factors exist that exert an influence on students' development so that any causal relationships cannot be inferred directly. One such example is students' chronological age: The students' age can have an impact both on assignment to grades and learning outcomes (Luyten and Veldkamp, 2011). In this way, the need for distinction between schooling and other factors that have been proposed to affect students' development in academic and non-academic outcomes arises in order to be able to make causal inferences on the effect of schooling.

The difficulty concerning the assessment of the contribution of education to students' development is that, since schooling is compulsory, almost everybody attends school. Even if there were a group of people who did not receive education, this group would not be representative of the characteristics of the population of interest: The universal nature of school attendance does not allow for the experimental investigation of the absolute effect of schooling (vs. no schooling) both in practical and ethical terms (Cahan and Davis, 1987; Luyten, 2006; Luyten and Veldkamp, 2011; Smith, 1972).

Several ways of overcoming this problem (e.g. value added and contextual value added models) have been proposed in the literature. One alternative is to use the difference between the learning rates during the school year and the summer learning rates (Heyns, 1978). In a review of 200 studies Ceci (1991) identified eight different designs that were used up to that time to assess schooling effects in non-experimental settings (Kyriakides and Luyten, 2009). The one that was characterised as the strongest was the RDD.

Figure 2.4: The Regression Discontinuity Approach (Luyten, Tymms and Jones, 2009)



To apply the RD approach to educational data it is necessary to have, at the very least, data on the attainment and the age of students from two adjacent grades (see Figure 2.4). Then, within each grade, the relationship between age and attainment can be estimated. This is usually obtained by the best fitting regression line of test scores on chronological age across the entire age range in that grade (Cook, Campell and Duay, 1979). It is expected that a discontinuity will be found among the oldest students in the lower grade and the youngest in the upper grade; this is depicted in Figure 2.4 by the vertical line connecting the mean achievement scores of the two ends of the age continuum. Assuming that the oldest students in the lower grade are relatively homogeneous in relation to characteristics other than age with the youngest students in the upper grade (Cahan and Cohen, 1989), and given that the age differences between students close to the cut off are negligible, this discontinuity can be interpreted as the effect of having

received an extra year of schooling (i.e. being in the upper grade). Note that this treatment effect is only identified locally at the cut-off – this can be considered as one limitation of the approach. In the RD framework, it is also possible to obtain an estimate of the net effect of one year's difference in chronological age in that grade: This is given by the slope of the regression line relating attainment and age within each grade (see Figure 2.4). When implemented within the multilevel modelling framework, the RD approach may also provide measures of the extent to which schools differ in their absolute schooling effects; thereby providing relative measures of effectiveness in addition to absolute measures of effectiveness.

Our knowledge of the absolute effect of schooling is limited, but available findings based on the RD approach (see, for example, Luyten, 2006; Cahan and Davis, 1987) indicate that more than fifty per cent of the progress that pupils make over a period of one year can be accounted for by schooling– over and above what can be explained in terms of student age and maturation). Moreover, evidence based on previous studies that used the RD approach suggests that one year of schooling does exert an effect and that this is about twice as strong as the effect of one chronological year (Cahan and Cohen, 1989; Cahan and Davis, 1987). More recent studies also confirm this result (Cliffordson, 2010; Cliffordson and Gustaffson, 2011; Luyten, 2006).

2.5.3 Modelling the Absolute Effect of Schooling

The conventional multilevel Regression Discontinuity model can be defined as follows:

$$y_{ij} = \gamma_{00} + \gamma_{10} age_{ij} + \gamma_{20} grade_{ij} + U_{0j} + U_{1j} grade_{ij} + \varepsilon_{ij} \quad (2.4)$$

In relationship (2.4) y_{ij} is the achievement of student i in school j while the variable age_{ij} is the age of the student - centered on the cut-off age (nine years for primary school data and thirteen years for secondary school data). The parameter γ_{10} represents the effect of one year of chronological age: it indicates the extent to which two students that are in the same grade but have one year's difference in their age, on average, in their achievement. Crucially, γ_{20}

represents the effect of grade level. In this way, γ_{20} is the overall "added-year" effect, the effect of one extra year of schooling, the "absolute schooling" effect. In the random component of the model U_{0j} represents the school-level residual that measures differences between schools in the achievement of their students in the lower grade, U_{1j} is the school-level residual that captures differences between schools in their absolute effect (of one year of extra schooling) and ε_{ij} is the student specific residual. The residuals U_{0j}, U_{1j} and ε_{ij} are assumed to be normally distributed with the school-level residuals (U_{0j}, U_{1j}) being independent from the student-level residual ε_{ij}.

2.5.4 The Effect of Age with the RD Approach

With the regression discontinuity model, it is possible to assess the effect of age on students' achievement - as well as the effect of one extra year of school. In the RD framework, the effect of chronological age is given by the slope of the regression line relating attainment and age within each grade (see Figure 2.4).The relationship between age and achievement in the lower and in the higher grade can be of particular interest, especially for educational systems that use a single age cut off criterion to allow student to begin school. Existing research suggests that the effects of age on achievement are more prevalent in early grades; they become weaker as children progress through school (Langer et al., 1984). However, recent studies in England (Crawford, Dearden and Greaves, 2011) suggest that the differences in outcomes between the youngest children in a year group – those born in August - and the oldest – born in September of the previous chronological year - can persist after secondary school and even later on, in students' working lives. This can be particularly worrisome in countries that base the assignment of students in the first year of primary school strictly based on the students' age. If the age differential between students in the lower grades, persist in later yearsthen this would have implications on the debates of whether or not such policies should be implemented.

2.5.5 Assumptions Underpinning the RD Approach

Strict adherence to the cut-off point is a critical requirement for the RD approach to yield accurate estimates: Admission to school should be based only on chronological age. Only then will the exact criteria that determine which students are placed in the lower and which in the upper grade will be known and, in this way, the researcher will be able to obtain unbiased effects of the "impact" of schooling (Shadish et al., 2002). Although some educational systems, for example schooling in England, meet this assumption, this is not always true (see also "The assumption of strict adherence to the cut off" in section 6.6.8 of "Discussion"). In some cases, this may be to some extent due to regional variations (e.g. in Australia and the United States). However, grade repetition or grade acceleration is also a common phenomenon in many educational systems that would severely compromise the application of the RD approach (Luyten et al., 2009; Luyten and Veldkamp, 2011). Importantly, this approach is not appropriate in systems in which there is considerable discretion in what age that students start school – based on parental choice, test scores or recommendations by pre-school teachers. Another basic assumption underlying the RD approach is that the student-level data across the two grades should be on the same scale. Moreover, correct modeling of the relationship between age and achievement is crucial (Luyten et al., 2009). If a linear function is estimated while, in fact, the function is quadratic or cubic, the RD estimates will be biased. For example, when the RD approach is applied to an extended range of analysis that spans over several years (see for example Kyriakides and Luyten, 2009), then there can be a deviation from linearity and a curve linear relationship may better characterize the data.

2.5.6 Advantages of the RD Approach over Conventional Approaches to School Accountability

The RD approach gives a very different perspective to school accountability compared to that of conventional value added modelling. The term "school effect", as defined in "Value Added Models of Educational Effectiveness and Relative School Effects" (see section 2.3.5), only

expresses the relative differences between schools in their effectiveness to raise the achievement of their students during a certain period of time. On the other hand, with the use of the RD approach, it is possible to obtain both estimates of one extra year of schooling (an estimate of the absolute schooling effect) and measures of the extent to which schools differ in their absolute schooling effects; thereby providing absolute in addition to relative measures of effectiveness.

The RD approach has been found to provide equivalent results both when applied to data collected in a cross sectional and also a longitudinal way (Luyten et al., 2009). Longitudinally, the same group of students should be followed for two consecutive years. In a cross sectional approach, the attainment of students of two equivalent groups of students in adjacent grades is required. This is an advantage over the traditional multilevel models that require longitudinal data of the attainment of students on at least two measurement occasions (Raudenbush, 1989; Guldemond and Bosker, 2009), since longitudinal data require more effort and money to collect.

Moreover, the RD approach provides a much stronger basis to conventional value added models for inferring causality. Causal inferences in value added models reflect, in fact, correlational relationships being based on assumptions that are often not met. The very assumption for causal inference, namely that no unmeasured variables exist that influence both selection into schools and achievement (Harris and McCaffrey, 2010; Rothstein, 2009) has been proved in many studies to be invalid because of pre-existing differences in student background variables that are not adequately adjusted for in value added models of educational effectiveness (see also section 2.3.8).

Importantly - and in relation to the methodological focus of my thesis, concerning bias that arises in estimation due to measurement error in the underlying data -- I note that in RD models, measurement error in achievement has less serious consequences for estimation than in value added modelling and conventional compositional analysis models. In RD models, individual achievement is modelled as a response variable (while in value added models prior

achievement is used as a predictor in the model). In this way, even if achievement is measured with certain amounts of measurement error, this has no serious implications for the estimates obtained with RD models (see section B.3 in Appendix B).

It should, nevertheless, be noted that there are certain limitations associated with the use of the RD approach for school accountability purposes (for an extensive discussion see section 6.6.8 of the literature review). For instance, the estimation of the relative school effects with RD approaches is based on observations on a limited number of students – those with age at the cut off. This can lead in larger standard errors in the estimation of the absolute effect of schooling for individual schools (added-year effects), so that relative differences across schools in their absolute effects are difficult to detect.

2.5.7 Investigating the Extent to which the Effect of one Extra Year of Schooling Correlates with the Effect of School Composition

The main focus of most of the studies which have applied the regression discontinuity approach in EER has been the investigation of the absolute effect of schooling and the extent to which schools differ in their absolute effects. Still, in a number of these studies there is some evidence on the extent to which variables relevant to the school's composition correlate with added-year effects. A methodological requirement for investigating this research question is to apply the RD approach in a multilevel modelling framework. Only then will the assessment of the relationships between variables that are situated at different levels be possible (see section 1.2).

Tymms, Merrell and Henderson (1997) applied the RD approach to data from the Performance Indicators at Primary School Project (PIPS) to explore the progress that the students made between the beginning and the end of the reception year. They concluded that schooling played an important role in knowledge. Moreover, they reported that pupils who attended a school with a high proportion of high able pupils tended to make more progress than those who attended a school where the general entry level was fairly low.

In a series of studies, Luyten and colleagues also used the RD framework to investigate the absolute effect of schooling with different datasets; in each of the studies implemented they also addressed the question on the extent to which the size of this effect correlated with school composition.

Starting with Luyten (2006), the researcher applied multilevel RD models to primary school mathematics and science TIMSS 1995 data He used data from eight different countries which participated in this assessment, including England – mathematics achievement data from the same database and from England were also used for the purposes of my thesis (see Study 3). He considered two school-level variables and the extent to which these correlated with added-year effects: the average number of books in a student's home and the gender composition of the school. He found no significant interaction for the school mean of books at home for either mathematics or science achievement. However, he suggested that the grade-level effect was stronger in girls' schools than in other types of schools.

Then the Luyten, Peschar and Coe (2008) study used PISA 2000 data on a sample of fifteen-year-old students in England. They assessed the absolute effect of schooling on reading performance, reading engagement and reading activities. A surprisingly modest effect was found on reading performance while none of the school-level variables considered in the study (the proportion of male students in the school, the proportion of students with a foreign home language and the school average parents' occupation) was found to have a significant effect on the absolute effect of schooling. Here I should, perhaps note that the data in PISA relate to older students than the data in TIMSS. This could be one explanation for the relatively small effects, since in general, the absolute effect of schooling is estimated to be smaller for older students.

In addition, the Luyten, Tymms and Jones (2009) study also applied the RD approach to investigate the absolute effect of schooling using PIPS data on students' achievement for mathematics reading and phonics. Although this study did not consider the potential effect of school composition effects on added-year outcomes, it should still be mentioned inasmuch as it demonstrated that the RD estimated obtained using cross-sectional data were not significantly

different from the RD findings produced using longitudinal data. This finding is especially important to my thesis in that I use cross-sectional data to assess the influence that one extra year of schooling exerts on students' outcomes (see section where I outline the data that I use in Study 3).

Heck and Moriyama (2010) was the first study, to my knowledge, to examine the usefulness of the RD approach, not only for defining the absolute schooling effect but, importantly, for determining whether between-school differences in the size of this effect could be explained by school-level factors. The researchers propose the implementation of the RD model in a multilevel structural equation modelling framework that allowed the assessment of the relationship between added-year effects with school leadership on school instructional practices, net of school composition and context. They defined school composition as denoting the student background and the teaching and administrative staff and school context as the school structure, size and facilities. Their results suggested that, net of context and composition factors, school leadership directly affected subsequent school instructional practices and, in turn, instructional practices affected added-year outcomes.

2.6 Concluding the Literature Review Chapter

A review of the Big-Fish-Little-Pond-Effect literature revealed that, establishing the generalizability of the phenomenon across different age groups can be of substantial value. Demonstrating the persistence of this phenomenon throughout several years of schooling is equally important. It is crucial to address relevant issues using the latest methodological developments stemming from self-concept research and the BFLPE paradigm: the Marsh, Lüdtke et al. (2009) multilevel structural equation modelling framework. In this way it is possible to compensate not only for the hierarchical structure that is typically prevalent in educational data (e.g. students nested within schools) but also for bias in the estimates due to measurement error at level 1 and level 2 and sampling error in the higher-level aggregates. In the second part of my literature review I have explained the way in which the inability of

multilevel models to take into account measurement and sampling error in the data results in biased estimates of compositional effects. This, in turn, may have serious substantive implications whenever relevant inferences are based on these – biased – estimates. Thus a more advanced framework is required which is able to correct for bias due to measurement and sampling error. A response to this methodological need in the existing literature is the Marsh, Lüdtke et al. (2009) framework. Interestingly enough, this framework is applicable not only in BFLPE models but, also in any research that relates to the investigation of the effects of the aggregate properties of a set of individuals on individual-level outcomes. An issue that has for some years received considerable attention in research into educational effectiveness is the extent to which the effect of school composition on students' outcomes can be effectively assessed. A review of relevant EER studies showed that school compositional effects, in the way in which they have conventionally been addressed seem to be inconsistent across different research settings (e.g. age of students involved in the sample, type of construct used to operationalize school compositional effects, models incorporated). One of the main reasons for inconsistent effects is the issue of level 1 under-representation in the way compositional models incorporated are specified. Two main facets of under-representation at level 1 have been identified: the prevalence of measurement error in the individual-level variable on which aggregation is based and omitted variables at level 1. In order to expand previous knowledge, I have suggested applying the 2x2 framework originally developed in educational psychology to EER research. This will solve potential bias in compositional effects estimates due to measurement error. In the last section of the literature review I introduced the RD approach and the way in which this can be applied in educational effectiveness research to investigate the absolute effect of schooling as well as differences across schools in their absolute effects. I ended this chapter with a review of a set of studies that have investigated the extent to which added-year effects are correlated with school composition.

Chapter 3: Phantom Effects in School Composition Research: Consequences of Failure to Control Biases Due to Measurement Error in Traditional Multilevel Models (Study 1)

3.1 Introduction

In educational settings, it is often of interest to investigate whether the "collective properties of a pupil body have an effect on an individual pupil's attainment over and above the effect of individual pupil characteristics" (Willms, 1985b, p. 33). Such effects are commonly referred to in the literature as compositional effects (e.g. Nash, 2003). Statistically, they are detected by aggregating student characteristics (e.g. prior achievement or socio-economic status) at the level of the school, or even at the level of the classroom or teacher, and investigating whether the obtained higher-level characteristic has an effect on the individual outcome variable, after controlling for the student-level variable on which aggregation is based (Lüdtke, Marsh, Robizsch, Trautwein, Asparouhov and Muthén, 2008).

The findings of studies investigating the effect of school compositional effects of average achievement on an individual student's achievement have not been consistent (see section 2.3.6 and section 2.3.7 in the literature review chapter for an extensive discussion of this issu, and so no consensus has been achieved on the magnitude and direction of this effect (Hattie, 2009; Wilkinson et al., 2000). With Study 1, I contribute to on-going debates on school compositional effects by proposing a methodological framework that addresses bias in estimation due to measurement error under-representation in compositional analysis models (see section 1.7 in the Introduction; see also section 2.3.8 in the literature review).

Study 1 consists of two distinct but interrelated analyses (denoted as Study 1a, Study 1b). The focus of both studies is on the compositional effect of school average mathematics achievement on students' progress in mathematics. The two studies have the same

methodological orientation, i.e. the way in which measurement error and sampling error adjustments alter estimates of school compositional effects. In Study 1a, I use year one and year four English primary school data. I adopt the 2 x 2 taxonomy of models proposed by Marsh, Lüdtke et al., (2009; see section 1.4; see also section 2.2.5 in the literature review), classified on the basis of adjustments made for measurement error and sampling error. I assess the effect of measurement error in the original data by fitting one model based on the total test scores, and another that treats items as multiple indicators. I adjust for sampling error by using latent aggregation rather than manifest aggregation of students' attainment to form the school-level variables.Study 1b is another application of the 2 x 2 taxonomy models. It incorporates year four mathematics achievement data from Cyprus to demonstrate how corrections for measurement error in themselves, or in combination with corrections for sampling error, alter compositional effect estimates.

In the subsequent sections I present the two studies that were carried out. The structure for each one is as follows: I first highlight the main aims of the study, outlining the research questions and research hypotheses addressed. Next there is an explanation of the methodology followed. Then the main findings are subsequently given and briefly discussed.

Phantom Compositional Effects due to Measurement Error Unreliability: Evidence from English Primary School Data (Study 1a)

My main concern in Study 1a was with the impact of the average achievement of the school's intake on the students' progress throughout the first four years of their primary education. Methodologically, I was concerned with the way in which the magnitude of this effect altered after adjustments for measurement error and sampling error in the data through the use of the four models of the Marsh, Lüdtke et al. (2009) 2x2 taxonomy (see Table 2.1). In pursuit of this aim, I used data from the Centre of Evaluation and Monitoring (CEM) in Durham and the Performance Indicators at Primary School (PIPS) test (www.pipsproject.org). The project is used throughout England and tracks aspects of schooling as pupils move through the primary sector, offering assessments of the children from the ages of three to eleven. For the purpose of my analysis, I used years one and four mathematics achievement measures. I assessed the magnitude and direction of the compositional effect of year one school average mathematics achievement on year four mathematics achievement after adjusting for year one individual prior achievement. Appropriate multilevel structural equation models from the Marsh et al. (2009) 2x2 taxonomy were applied. I showed how the parameter estimates and associated conclusions changed when adjustments were made for either or both (i) measurement error at level 1 and level 2 and/or (ii) sampling error.

Subsequently, I performed a small simulation study, following Harker and Tymms (2004) in their classic demonstration of "phantom compositional effects" (see section 1.7). In this way, I investigated how the estimates of the doubly manifest compositional models would change had the baseline assessment been less reliable. Of special interest was the direction of the bias in the compositional effect as random error of increasing variability was deliberately added to the level 1 variable. I extended Harker and Tymms's (2004) study by showing how

relevant models from the 2x2 taxonomy, namely the latent manifest and the doubly latent model (see Table 2.1 in which the four models of the 2x2 taxonomy are given), could compensate for the bias produced by the reduction of measurement error reliability. I demonstrated that these approaches could retrieve the estimates of the original analysis even for datasets with lower measurement error reliability.

In addition, as well as investigating the way in which the estimated within-group effect and the compositional effect changed as random error was deliberately added to the baseline assessment, I observed the impact of random error on random effects estimates (residual variance at the level of the student and at the level of the school) in compositional models (see section 2.2.1 for the equation desciring the different components of a compositional model). I investigated the extent to which the use of multilevel structural equation models from the 2x2 taxonomy could also correct for bias in the random effects estimates of compositional models – as well as for bias in the fixed effects estimates.

Random effects in compositional models are of special relevance in studies of school accountability and value added analysis where they are used to assess the magnitude and direction of school effects (see section 2.3.5 where I explain the use of value added models and compositional effects in school effectiveness research; see also section A.1.3 in Appendix A). In fact, the compositional models invorporated in Study 1a, described the simplest form of a "contextual value added model" with only prior achievement and average prior achievement controlled for. (see section 2.3.5 in the literature review for a definition of the contextual value added model; see also section A.1.1 in Appendix A for the statistical equation describing a value added model in which compositional effects are adjusted for). Since no extensive controls for student background variables were made in the models used in Study 1a, the school effects estimates derived in my analysis were not of interest in themselves; rather the focus was on comparing the school effects in the students' mathematics achievement (i) across datasets with

different measurement error reliability at level 1 (ii) across the four models of the 2x2 taxonomy.

3.2 Research Questions and Research Hypotheses for Study 1a

My focus of enquiry in Study 1a lay mainly on the effect of school average mathematics achievement in year one (L2-MACH1) on individual mathematics achievement in year four (L1-MACH4), over and above the effect of individual mathematics achievement in year one (L1-MACH1).

I begin this chapter with reference to the conventional approach to compositional analysis, referred to in the 2x2 taxonomy as the doubly manifest approach (see section 2.2.5 and Table 2.1 in the literature review where the four models of the 2x2 taxonomy are described). The model is applied initially to the original data and then to the simulated data, investigating how the within-group effect and the compositional effect alter as the baseline measure becomes less reliable. This first part of Study 1a is a replication of an analogous simulation study that was conducted by Harker and Tymms (2004) and demonstrates how phantom compositional effects can arise simply due to the prevalence of random measurement error in the baseline measure on which aggregation is based.

I then proceed with the research questions relevant to the application of the partial and full correction approaches from the 2x2 taxonomy that make adjustments for measurement error. To correct for the bias due to measurement error unreliability in the baseline scores, I employed the latent manifest and the doubly latent approach (see Table 2.1 in which I give the four models of the 2x2 taxonomy), one at a time. They both make corrections for measurement error bias through the use of multiple indicators. Their difference lies in the fact that the doubly latent approach makes additional adjustments for sampling error assuming latent aggregation to form the school-level constructs (see section 2.2.5 in the literature review in which the four models of the 2x2 taxonomy are described). The focus is on how the estimates of the

compositional effect and the within group effect change when different adjustments are made in the data for measurement error alone or in combination with sampling error. Based on theoretical derivations and empirical results (see Ferrão and Goldstein, 2009; Fletcher, 2012; Marsh, Lüdtke, et al., 2009; Woodhouse et al., 1996) hypotheses can be made in relation to how the estimates should alter after such corrections.

One crucial enquiry concerns the extent to which the models from the 2x2 taxonomy can correct for measurement error bias even for data with substantially lower reliability. This is investigated by applying the partial and full correction approaches to the simulated data in which the reliability of the baseline is deliberately made lower.

Although in my analysis my focus was mainly on the effect of individual mathematics achievement and the compositional effect of school average achievement, the way the estimates of the random effects changed across the four models of the 2x2 taxonomy in this first analysis was also investigated. In the last paragraph of the present section, I form relevant research hypotheses based on previous studies.

3.2.1 Applying the Doubly Manifest Approach to the Original Data

Much of the research on the compositional effects of school average prior achievement on subsequent achievement is weak with conflicting results (see sections 2.3.6, 2.3.7 and 2.3.8 in the literature review chapter), largely suggesting weak positive effects. However, most of this research – even sophisticated value added studies -- is likely to suffer from phantom effects so that there is no clear basis for predicting even the direction of compositional effects.

Research Hypothesis 1a.1: The effect of individual mathematics attainment in year one (L1-MACH1) on subsequent mathematics achievement (L1-MACH4) is positive and significant.

Research Question 1a.2: What is the size and the direction of the effect of school average attainment in year one (L2-MACH1) of primary school on attainment in year four of primary school (L1-MACH4)?

3.2.2 Applying the Doubly Manifest Approach to the Simulated Data: Demonstrating Measurement Error Bias

Measurement error in the individual-level predictor leads to a negative bias in the estimated within group effect and, since the effect of prior level 1 achievement on subsequent level 1 achievement is typically positive, it is expected that the estimate of this effect will be attenuated (i.e. estimated less positive) as a result of the bias introduced by measurement error at level 1. At the same time measurement error in the individual-level measures leads to a positive bias in the estimated compositional effect of the corresponding aggregate, to phantom compositional effects. This has been shown, for instance, by Harker and Tymms (2004; see section 1.7 in the introduction and section 2.3.4 in the literature review where I refer to the study conducted by Harker and Tymms). I also refer the reader to section 2.2.3 of the literature review in which I discuss more extensively the impact of level 1 measurement error on the estimates of within-group effects and compositional effects in compositional analysis.

Research Hypothesis 1a.3: When simulated random measurement error is added to the baseline year one assessment, a negative bias occurs in the estimated within-group effect while a positive bias is prevalent in the estimated compositional effect. This bias becomes larger the larger as the amount of measurement error increases.

3.2.3 Correcting for Bias due to Measurement Error Unreliability

To correct for the positive bias in the estimates of the compositional effects due to measurement error unreliability in the baseline scores, I employed the latent manifest and the doubly latent approach (see Table 2.1), one at a time. While they both make corrections for measurement

error bias through the use of multiple indicators, the difference between them lies in the fact that the doubly latent approach makes additional adjustments for sampling error assuming latent aggregation to form the school-level constructs (see section 2.2.5 in which I discuss the four models of the 2x2 taxonomy). In section 2.2.3 of the literature review concerning measurement error bias in compositional analysis estimates (section 2.2), I discussed extensively the potential consequences that measurement error can have on the estimates on compositional analysis. I form Research Hypothesis 1a.4 – Research Hypothesis 1a.6 based on this discussion.

Correcting for measurement error in the individual-level variable will always lead to dis-attenuated estimates of the level 1 effect. The models of the 2x2 taxonomy (the latent manifest and the doubly latent approach) make adjustments not only for measurement error in the individual-level variable, but also in the aggregated variable. It has been shown, nonetheless, that measurement error at level 2 does not really have any impact on estimation at level 1 (see Woodhouse, Yang, Goldstein and Rasbash, 1996). Therefore the following hypothesis can be made:

Research hypothesis 1a.4: Adjusting for measurement error using the models of the 2x2 taxonomy (the latent manifest and the doubly latent approach) will dis-attenuate the within-group effect of prior achievement on subsequent achievement.

In the same way, adjustments for level 1 measurement error should correct for the positive bias in the estimated compositional effects. Of interest is the impact of simultaneous adjustments for measurement error at both level 1 and level 2 through the models of the 2x2 taxonomy. Generally, the impact of level 2 measurement error adjustments on level 2 effects should not be as severe as the impact for level 1 measurement error adjustments (see for example Goldstein, Kounali and Robinson, 2008). The following research question is posed:

Research question 1a.5: What is the impact of adjusting for measurement error at level 1 and level 2 using the models of the 2x2 taxonomy on the estimated compositional effects of school average mathematics achievement?

It is expected that the latent manifest and the doubly latent approach will correct for measurement error bias in not only the original data but also in simulated data:

Research hypothesis 1a.6: The application of the latent manifest and the doubly latent approach eliminates measurement error bias even when measurement error unreliability is substantially low.

3.2.4 Corrections for Sampling Error

Research question 1a.7: How does adjusting for sampling error through the models of the 2x2 taxonomy (manifest-latent and doubly latent approach) affect the estimated within-group effect and compositional effect?

The direction of bias in the compositional effects due to sampling error depends, amongst other factors, on whether the actual effect is positive or negative. Based on mathematical derivations, Lüdtke et al. (2008) suggest that when adjustments are not made for sampling error, compositional effects are underestimated (negative effects are estimated less negative and positive effects are estimated less positive). In other words, the relation between the group-level construct and the individual outcome is estimated to be weaker when sampling error is prevalent in the data. This bias is larger for smaller group size and for a lower intra class correlation

(ICC ; see sectionA.1.2 in Appendix A), suggesting less agreement in the observations within the same group (see A.2.3 in Appendix A).

3.2.5 The Impact of Measurement and Sampling Error Adjustments on Standard Errors

Estimates of statistical parameters are always associated with a standard error. This quantifies the standard deviation of the distribution of all the potential values that the estimate can take using all possible samples of the same size from the population of interest. Standard errors are indicative of the accuracy with which estimation is made. Using simulated data Lüdtke, Marsh, Robitzsch and Trautwein (2011) have shown that while the partial and full correction approaches from the 2x2 taxonomy reduce the bias in the parameter estimates, they introduce variability in resulting estimates. The quite complex doubly latent approach in particular may result in too large standard estimates of the when simultaneously correcting for two sources of error**.** In contrast, the doubly manifest approach and the latent measurement/manifest aggregation approach that do not make adjustments for sampling error tend to underestimate the sampling variability of the between group coefficient (leading to smaller standard errors).

Therefore, I form the following research hypothesis:

Research hypothesis 1a.8: Although the partial and full correction approaches correct for bias due to either or both measurement and sampling error, they will generally give larger standard errors of the estimated compositional effects than the doubly manifest approach.

3.2.6 The Impact of Measurement Error on Random Effects Estimates

In this section I make predictions on how the estimated random effects (residual variance at level 1 and level 2) in a compositional model (see section A.1.1 in the literature review where I provide this model) change as the baseline measure becomes less reliable. I focus on the extent to which adjustments for measurement error through the models of the 2x2 taxonomy correct for relevant bias and retrieve the random effect estimates of the original analysis.

The substantive importance of the investigation of random effects in compositional models has to do with how they are used in value added models of educational effectiveness to quantify relative school effects. School effects in value added models of educational effectiveness are defined by the ratio of the school-level residual variance divided by the total variance in students' achievement. Measurement error unreliability in the variables involved in value added models (e.g. prior achievement, school-average prior achievement as in my analysis) can have consequences on conclusions on the size and the magnitude of school effects estimates. These are also examined as part of Study 1a.

Research Question 1a.9: How are the estimates of the random effects at level 1 and level 2 and, similarly, the estimates of school effects altered as the baseline assessment becomes less reliable?

The role of measurement error at level 1 and level 2 on the random effects of value added models has been addressed to a certain extent by Woodhouse et al. (1996). The researchers explain that adjustments for measurement error at level 1 can cause smaller level 1 residual variance but almost unchanged level 2 residual variance. In this way the estimated school effects (proportion of the total variance in the outcome attributed at level 2) become larger once measurement error is adjusted for. Moreover, they show that additional adjustments for measurement error at level 2 lead to smaller level 2 residual estimates. But this change is so small compared to the reduction in the level 1 variance that the school effects are again estimated to be larger. The effects of measurement error on the residual terms should be more prevalent when no adjustments are made for the school-level aggregate. When compositional effects are included in the models, they can compensate for level 1 measurement error (off-setting biases) so that the residuals are not much affected even though the level 1 and level 2 estimated effects may be biased by measurement error.

Research Hypothesis 1a.10: Adjustments for measurement error lead to larger estimates of the school effects.

Research Question 1a.11**:** What is the impact of adjustments for sampling error on the estimates of school effects?

3.2.7 Summary of Main Research Questions and Research Hypotheses of Study 1a

Study 1a incorporates primary school data from a large sample of students in primary schools in England obtained from the Performance Indicators in Primary School (PIPS) project. The study employs a simple form of a contextual value added model to evaluate the compositional effect of school average mathematics achievement at the end of year one on year four mathematics achievement, after adjustments for individual achievement at year one. I evaluate the extent to which random measurement error in the baseline score is a source of bias in the estimated within-group and compositional effect obtained by the conventional approach to compositional analysis. I also test the impact of random measurement error at level-1 on random effects estimates. These are used to assess the magnitude of the school effect estimates in contextual value added models used for the purposes of school accountability. Importantly, I propose how bias due to measurement error and/or sampling error through the Marsh, Lüdtke, et al.(2009) 2x2 taxonomy of models can be corrected for.

Having outlined the main aims of Study 1a I proceed with the methodology followed to provide answers to my research questions and research hypotheses.

3.3 Methodology for Study 1a

A description follows of the methodology used to answer the research questions and research hypotheses that I have outlined. I begin with a reference to the centre for Curriculum, Evaluation and Monitoring and the various projects it runs; the focus here is on the Performance Indicators at Primary school tests – this is where the data for Study 1a were obtained. Then follows a description of measures, as provided by this assessment and the data samples involved in the analysis. An extensive discussion on missing data analysis follows. The intention of the last part of the present chapter is to clarify the exact way in which the variables necessary for the analysis were calculated based on the available data and to emphasize the differences between the four models of the 2x2 taxonomy.

3.3.1 Performance Indicators at Primary School

For the purposes of my research, I used data obtained by the Performance Indicators at Primary School (PIPS) monitoring project (www.pipsproject.org), run by the Curriculum, Evaluation and Monitoring (CEM) centre in Durham University. Data on pupils' attainment have been collected since 1993 as part of this system (Tymms, Jones, Albone and Henderson, 2009). The PIPS tests track aspects of schooling as pupils move through the primary sector, from the age of three until their eleventh year of age. The tests are administered as traditional paper and pencil tests but they are also available to be computer delivered. They include three sections: one that measures the mathematics abilities of students, one that measures their reading ability and another that provides a measure of the child's "developed" ability. The last is formed by combining two curriculum independent measures of vocabulary and non-verbal ability. Attitudes towards mathematics, reading and school are also assessed. In the last year of primary school, there is an extra assessment in Science.

Apart from PIPS, the CEM centre has developed a number of other widely used monitoring systems. These include the Advanced Level Information System (ALIS), the Year 11 Information System (Yellis) and the Middle Years Information Systems (MidYIS). Although the data used in the present analysis related to England (Jones, personal communication, 2010), the work of the centre is also used throughout Wales and Northern Ireland. Some of its work is even run internationally (Tymms and Coe, 2003). For example, a CEM centre is well established at Christchurch in New Zealand and runs a number of projects like YELLIS, MidYIS and the PIPS on-entry baseline, a version of PIPS. The parallel version to ALIS is BLIS (the Bursary Level Information System). The materials are also used in Australia and Hong Kong.

The educational institutions themselves pay for the assessments and feedback. Therefore not all educational institutions participate in the program and not all schools use PIPS assessments at the end of the years of the primary school. Nevertheless, previous analyses have indicated that all the samples used in the various projects of the CEM centre, except the one from the pre – school, are representative of England as a whole (Tymms, Merrell , Heron, Jones, Albone and Henderson, 2008).

3.3.2 Measures and Data Samples

The data used consisted of 19,059 students from 593 schools, were longitudinal in nature and were collected for the same students in years both one and four. The students involved entered primary school in the academic year 2004-2005. The PIPS year one and year four mathematics tests used for the purposes of Study 1a were based on item-level data. Each item in the assessment was given a value of one if it was answered correctly, a value of zero if it was answered wrong, and was left blank if it was not answered at all. The way in which these item-level data were used to construct the variables that I incorporated in my analysis is described in a following section (see section 3.3.4 on "Variables").

The data used in my study were chosen strategically from a larger multi-cohort multi-wave database, part of the PIPS project provided by colleagues in the CEM centre. Five different sets of students were present in the database, the oldest cohort entering primary school in the academic year 2000-2001 and the youngest in 2004-2005. Although there were available data for six years of primary school with each student participating in at least two year tests, I chose to include data on only year one and year four in my analysis. This choice was taken partly due to technical issues related to the quality of the data available for the different years – for example the data for year one and for year four were more complete than data for other years. Most importantly, I found it intriguing to investigate the effect of school prior achievement on individual achievement somewhere halfway through the schooling experience of students in primary school. Studies on the impact of average prior attainment on student achievement usually focus on the last year of primary school – year six (e.g. Bondi, 1991). The magnitude of such effects on achievement in earlier years has been only sparsely assessed (e.g Lauder, Kounali, Robinson, Goldstein and Thrupp, 2007; Strand, 1997).

From among the total number of schools involved in the selected sample, I based my analysis only on information from schools that participated in both year one and year four educational assessments. These were identified as those with at least one student with data on both measurement occasions (year one and year four) – although most schools had many students (see Table 3.1). From these schools I selected those students who took the PIPS test in year one or year four, even if they had missing data in one of these two years**.** I focused on the youngest group of students (entering primary school in the academic year 2004-2005); this comprised those students who took either their year one assessment in 2005 or their year four assessment in 2008.

3.3.2.1 School identification

The models that I incorporated (see section 2.2.5; see also section A.3 in Appendix A) required the use of single school identification for each student – this was used as the level 2 unit for the purposes of the two-level analysis. The specification of this unit was somewhat problematic for:

- Students who did not participate in year one or in year four assessment, in that the school id was not available for these students. For these cases, the school reported for the single year in which they took a test, either year one or year four, this being specified as the level 2 unit in my analysis.

- Students who were not in the same school in years one and four. The problem of students changing school during the first four years of their primary education relates to the extent to which the year one average achievement was a valid index of the school composition in our analyses. If a significant proportion of students had changed school, the year one average achievement could not really capture the composition of the school accurately. Student mobility was not particularly problematic for our analysis. Among students who took both tests, 17,681 (92.8%) remained in the same school for all the first four years of their primary education while 1,281 (6.7%) changed school once and 97 (.5%) changed schools more than once. For those who changed school, if their year four school was among those participating in both year one and year four assessments, this was specified as the level 2 unit. Otherwise their year one school was specified as the level 2 unit.

3.3.2.2 Delayed/ accelerated students

Among the total number of students involved in my analysis,(19059) there were five cases who, despite having had their initial assessment at a year earlier than their peers (2002 or 2004), actually received their year four exams in year 2008. These were students belonging to older

groups that were, nevertheless, tested in year four with younger students. The interpretation was that these students were those who were retained for one or more years (remained in the same class for one or more years). As their percentage of the total sample was negligible, they were kept in the analysis. Moreover, fifteen students (0.1%) in the sample seemed to have taken their year four examination a year earlier than 2008 although they began primary school in 2004 and were, therefore, classified in the youngest group of students. These could have been those accelerated during their first years in the school. Although for most misclassified cases, the reported school year could be interpreted as either due to a delayed or accelerated schooling career, for three students a completely wrong code appeared as the year in which they took their year four assessment. This was replaced by 2008.

3.3.2.3 Numbers of schools and students

In the final dataset, there were 19059 students in 593 schools with an average of 40 students per school. In Table 3.1, the frequencies for the schools' sample size are indicated.

Table 3.1: Number of students sampled from within each school in the dataset used in Study 1a

Number of students	**Number of schools**	**Proportion of schools**
40≥	153	.258
38≤ and <40	17	.030
36≤and <38	16	.027
34≤and <36	26	.044
20 ≤ and< 34	266	.449
20	22	.037
8 ≤ and< 20	129	.218
≤8	20	.034

Note. The range of the number of students within schools in my analysis is displayed in the first column. The number of schools with sample size in the range given in the column is given in the second column and the corresponding proportion out of the total number of schools participating in the analysis in the third column.

3.3.2.4 Use of two different versions of the same test

For both the year one and year four assessment, there were two different versions of the same test, version A and version B. Basic descriptive analysis showed that about half students in the sample took version A and the other half took version B. Both versions had identical questions in the same order, but the order of the response options was different; this facilitated practical aspects in relation to the administration of the tests (Coe, personal communication, 2010). In my analysis, I used all the assessment results and treated them in the same way independently of which version of the test was taken. Descriptive statistics for the measures incorporated were performed for each version and the means were compared to each other using independent t-tests. Generally, there were no significant differences observed across the two versions.

3.3.3 Missing Data

In this section, I distinguish between two types of missing data. A unit non-response is a term used for cases who did not participate in the PIPS assessment at all in one particular year. Item non-response refers to cases for which only partial data were available in that a person participated in the test but did not respond to certain individual items (see also Schafer and Graham, 2002).

3.3.3.1 Defining unit non-response

Missing units in the analysis were identified as those not having a matching assessment in the corresponding year. In this way, from the total number of 19059 students in my sample there were 2085 (10.9%) students who did not participate in the PIPS end of year one mathematics test and a slightly smaller proportion (1707, 9%) who did not take the year four mathematics assessment. For the mathematics achievement section, missing units were defined not only as those students who did not participate in the assessment, but also (i) those who did not complete the mathematics section, even though they participated in the assessment and (ii) those who completed only an inadequate number of items in the mathematics section, which did not allow for reliable inferences to be made in relation to their mathematics ability. For both year one and year four, I considered the minimum number of items in the mathematics achievement section that a student should have completed before being included in the analysis as a non-missing case. I derived the percentage of cases that answered at least one and at least five items (see Table 4.2). Using the "more than one" rule increased the number of missing cases to 2085 (10.9%) for year one and 1707 (9%) for year four. Using the "more than five" rule resulted in datasets with an even larger number of missing cases (see Table 3.2) – 2289 (12%) for year one and 1772 (9%) for year four. Any case with five or fewer items attempted in the test was treated as a unit non-response even if a matching code for the PIPS assessment was available for this case.

3.3.3.2 Treatment of item non-response for mathematics achievement data

For students with at least five completed items in the mathematics achievement section (non-missing units), values were given to the unanswered items based on the probability of guessing the item correctly. For example, for a multiple choice question with four options, a missing value was replaced with .25. For an item with an open question, a missing value was assigned a value of zero. Some alternative ways to treat item non-response were also considered in the analysis (see supplementary analyses to Study 1a as given in section A.4); the results obtained from these analyses did not substantially differ from each other.

Table 3.2: The numbers of non-missing for each year and the number of students who attempted more than one and more than five items respectively for the PIPS mathematics achievement tests

Year group	**valid cases (% of total cases)**	**missing cases (% of total cases)**	**students with more than one item (% of valid cases)**	**students with more than five items (% of valid cases)**	**total number of cases**
Year one	16974 (89.1%)	2085 (10.9%)	16915 (88.75%)	16770 (88%)	19059
Year four	17352 (91%)	1707 (9%)	17293 (90.7%)	17287 (90.7%)	

Note. Among the total number of students (19059) some participated in the PIPS Year one test and some in the PIPS Year four test. In the first and second columns of the table the total numbers of valid and missing cases are given. Among the valid cases I distinguish those who attempted more than one item, and thus took the mathematics section of the PIPS test (third column), and those that appeared to have attempted more than five items (the fourth column). These numbers are the ones representing my data sample.

3.3.3.3 Use of item parcels

Parceling is a commonly used measurement practice with latent variable analysis techniques (Little, Cunningham, Shahar, Widaman, 2002). It involves the use of parcels –aggregate-level indicators comprised of the average of groups of items. In this way, the number of indicators associated with each construct is substantially reduced resulting in a more parsimonious model

with fewer parameters to be estimated (Marsh, Lüdtke, Nagengast, Morin and Von Davier, 2013). I also note that because the original items were dichotomous, the use of item parcels gave indicators with a distribution closer to normal – facilitating normal theory-based estimation (Marsh et al., 2013).

An important assumption in the use of item parcels is that the items form a uni-dimensional construct (see Little et al., 2002). In preliminary factor analyses I ascertained that the set of items designed to measure a single factor of math achievement were reasonably unidimensional both for year one achievement data and for year four achievement data.

I used item parcelling to form the multiple indicators for mathematics achievement at year one and year four. I formed the parcels after item non-response was treated as explained in the previous paragraph. For year one, I formed three nine-item parcels, taking the average of every third item available for year one assessment; there were twenty seven items altogether. In the same way, for year four I formed four nine-item parcels (there were thirty six items altogether). The distribution of the items across the parcels was performed in such a way that the mean value for each parcel as well as its standard deviation was approximately the same. The single-level indicators for mathematics achievement at year one and year four were formed by taking the average score across the parcels for each year – equivalently by taking the average score over all the items for each year. In this way, three indicators were derived for the mathematics achievement of students in year one and four indicators for the mathematics achievement of students in year four.

3.3.3.4 The use of multiple imputation to treat missing data

Multiple Imputation was used to treat unit non-response for the mathematics achievement data. The procedure generally involves replacing missing values with a list of two or more simulated values. In this way, plausible alternative versions of the complete data are produced. Each of

these is analysed by a complete-data method. Then the results from each imputed dataset are combined to obtain overall estimates and standard errors.

Multiple imputation (MI), first introduced in the late 1980s by Rubin (Rubin, 1987), has been characterized as the "practical state of the art" (Schafer and Graham, 2002, p. 173) among likelihood procedures (Maximum Likelihood Estimation) to treat missing data. It has been highly recommended by researchers such as Rubin (1996) and Schafer and Graham (2002) as the "method of choice" (Rubin, 1996, p.473) for addressing problems due to missing values. Alternative approaches include list-wise deletion and ad-hoc methods that replace missing values in the dataset with a fixed estimate of the missing data (e.g. mean substitution). Nevertheless, the use of such methods has been strongly criticized (e.g. Wilkinson and Task Force on statistical inference, 1999; see also Xu, 2010) as being statistically invalid for scientific estimates. Other commonly used (single) imputation methods (e.g. ratio or regression imputation) can also lead to wrong inferences and, in particular, to serious underestimation of the true variance in the imputed estimator because unreliability due to unknown missing values is not taken into account. The standard errors in MI compensate for this in that they reflect missing data uncertainty as well as finite sample variation.

MI rests on the assumption of data "Missing at Random" (MAR) as defined by Rubin (1976). This means that the distribution of the missing data should not depend on the missing variables, only on the observed data. A special case of MAR occurs when the missing data depend on neither the observed nor the missing data (data "missing completely at random"-MCAR). In general, it is not possible to test whether MAR holds in a dataset. It could be the case, for example, that the missing data are "missing not at random" (MNAR); the probability of missingness might depend on missing data as well as on observed variables. However, Collins, Schafer and Kam (2001) have demonstrated that an erroneous assumption of MAR has only a small impact on estimation using MI (see also Schafer and Graham, 2002).

For the purposes of my study I used SPSS to impute missing data values. The imputation method followed was fully conditional specification. This is an iterative Markov chain Monte Carlo (MCMC) method that involves a specification of a group of variables to be used in the imputation model –these comprise the variable list – and a specification of a number of iterations that should be performed before obtaining the imputed values. In each iteration and for each variable in the order specified in the variable list, it fits a univariate model using the variable to be imputed as a dependent variable and all the other variables in the model as predictors. It subsequently imputes missing variables for the variable being fitted. The method repeats the procedure until the specified number of iteration is reached and the imputed values of the final iteration are saved to the imputed dataset.

In specifying the variables to be included in the imputation model I chose to include the three parcels corresponding to achievement in year one and the four parcels related to achievement in year four. In the imputation procedure, I also considered information on other variables available in the data. Specifically, the scores of students in other tests provided in the PIPS assessment (see section 3.3.1) as well as measures of students' attitudes towards reading and towards school, also provided in the database were included in the variable list and were defined as predictors in the imputation model.

3.3.3.5 Taking the multilevel structure into account

I used a two-stage imputation procedure to allow for the multilevel structure of the data. In the first stage, a single imputed dataset was produced in which the missing values of the student-level variables denoted in the variable list were imputed and appropriate student-level variables were used as predictors (see previous paragraph). I remind the reader that the variables to be imputed were the parcels for year one and year four mathematics achievement. Then the corresponding school-level aggregates (school-level average) were computed based on the complete data. These were merged back to the initial student-level dataset and a second set of

imputations was performed. The set of variables to be imputed was the same as in the first imputation. In this way, not only student-level variables were considered as predictors in the imputation model, but also school-level predictors; those obtained by aggregating student-level data obtained by a single imputation.

3.3.3.6 Maximum number of iterations and number of produced datasets

Based on recommendations by Rubin (1987) five imputed datasets were generated. Rubin suggests that the efficiency of estimation based on a finite number of imputations, say μ, relative to one based on an infinite number is $(1+\lambda/\mu)^{-1}$ where λ is the rate of missing information. Replacing μ with 5 and λ with .1 (a value that approximates the rate of unit-missing data in my analysis; see Table 3.2) gives an estimate of .98 – high enough for the purposes of my analysis.

3.3.4 Variables

In the imputed datasets the data on students' mathematics achievement were in the form of multiple indicators measuring student-level achievement. These could be used for the application of the latent manifest and the doubly latent approach that make multivariate adjustments for measurement error (see section 2.2.5, in which I describe the four models of the 2x2 taxonomy and section A.3 Appendix A in which I provide a detailed description of the models).

3.3.4.1 Student-level variables

The models that were manifest in terms of measurement error (the doubly manifest and the manifest latent approach; see also Table 2.1 in section 2.2.5) required only the use of single scale scores for each of the students' mathematics achievements in year one and year four. For each dataset, this was obtained by taking the average of the imputed parcels – the multiple

indicators for the mathematics achievement of students. For models that control for measurement error (the latent manifest and the doubly latent approach) the parcels themselves were used as multiple indicators from which the value of the latent construct was to be inferred.

3.3.4.2 School-level variables

For the doubly manifest approach, the school-level variable for mathematics achievement in year one was formed by taking the average of the individual-level mathematics achievement in year one across all students in the school. For the manifest latent approach, latent aggregation was assumed, using a single indicator for the latent school average mathematics achievement at year one – the observed average mathematics achievement at year one. For the latent manifest approach multiple indicators were used to adjust for measurement error in the school-level achievement and each was formed by taking the average of the corresponding individual-level indicator across all students in the school (i.e., the aggregation was manifest and did not take sampling error into account). Lastly, for the doubly latent approach, latent aggregation was followed to obtain the school-level indicators from the individual-level indicator.

3.3.4.3 Standardisation

All the variables at level 1 were standardized in relation to the total sample by subtracting the overall mean and dividing by the overall standard deviation, so that they had a mean of zero and a standard deviation of one. The school-level aggregates were based on the standardised scores but were not re-standardised at level 2 so that level 1 and level 2 were on the same metric.

3.3.5 Statistical Analysis

I initially fitted the four models of the 2 x 2 taxonomy to the data, regressing year four individual mathematics achievement on year one mathematics achievement and the corresponding group-level aggregate. All four models were specified in Mplus 6, assuming group mean centring (Kreft, de Leeuw and Aiken, 1995). This implied that the regression

coefficient of the school-level aggregate on individual achievement was not a direct estimate of the compositional effect; instead, the within-group effect should be subtracted from the between-group coefficient to obtain this estimate (see also Appendix A.1.1). Measurement invariance of the latent outcome variable (mathematics achievement in year four) across the levels of student and school (Marsh, Muthén et al., 2009) was thus established, to facilitate the estimation of the compositional effect, as the difference between the between-group effect and the within-group effect (see section 2.2.1).

A small simulation study was also conducted, with a two-fold purpose: (a) to investigate the direction of bias in the compositional effect for increasingly lower reliability estimates of the baseline measure, and (b) to examine the behaviour of the models of the 2 x 2 taxonomy, and in particular of the latent manifest and the doubly latent approach, which make adjustments for measurement error, for data with substantially lower measurement error reliability than the original data.Based on the squared standardised factor loadings of each of the three multiple indicators that were formed for year one mathematics achievement (see section 3.3.4 in which I describe the variables that were used as part of my study), the overall reliability of the observed score was .875. Importantly, for the purposes of our investigation, the original data were highly reliable. Increasingly large amounts of random error were added, so that the overall reliability was reduced to .75 and .56 (including measurement error already in the data). The conventional doubly manifest approach was applied to the original dataset and to each of the two datasets with added simulated error. The focus was on how the estimated compositional effect changed across the three datasets. Because of the way the reliability was manipulated, any change in the observed compositional effect across the three analyses could only be attributed to the added measurement error. We then used the latent manifest approach and the doubly latent approach to correct for measurement error bias in estimation. Each of these makes different assumptions for sampling error (see Table 2.1, where I give the four models of the 2x2 taxonomy); this was taken into account in the interpretation of the findings (see section 3.4 on the results for Study 1a).

3.3.6 Calculation of the effect Size

One of the main purposes of my analysis was to compare the estimates given by the four different models with each other. To achieve comparable results, I used effect size to assess the magnitude of the effect of individual and school average ability on year one on individual attainment at year four. The effect sizes were given by the following equations:

$$ES_{\beta_{com}} = 2 * \beta_{com} * SD_{com} / SD_{outcome} \quad (3.1)$$

$$ES_{\beta_{within}} = 2 * \beta_{within} * SD_{within} / SD_{outcome} \quad (3.2)$$

Equation is used to calculate the effect size of the compositional effect ($ES_{\beta_{com}}$) and equation is used to calculate the effect size of the individual effect ($ES_{\beta_{within}}$). The denominator for both equations is the same, the standard deviation of the observed score of the students in year four mathematics score. In both equations the unstandardised regression coefficients, β_{com} and β_{within} respectively, are multiplied by the standard deviation of the individual-level predictor. This effect size is comparable to Cohen's d (Nagengast and Marsh, 2011).

3.3.7 Evaluation of Model Fit

An important caveat in the use of Structural Equation Modelling is the correct assessment of the goodness of fit. Appropriate tests should be performed by applied researchers in order to evaluate the extent to which the imposed model can be reasonably well supported by the data. For the purpose of my study I used a range of different fit indices, following the recommendations of researchers such as Marsh, Balla and Hau (1996) and Jaccard and Wan (1996). I use measures of empirical discrepancy (Streiger and Lind, 1980), such as the chi-square (χ^2) and the root mean square error of approximation (RMSEA), and measures from the family of incremental fit indices (Hu and Bentler, 1999), namely the Tucker – Lewis Index

(TLI) and the Comparative Fit Index (CFI). The former class of statistics represents the difference between the fitted covariance matrix, that predicted by the model, and the sample covariance matrix. The latter (incremental fit indices) compare the model of interest with some alternative, such as the null or the independence model.

Although indices of fit do help in the evaluation of the model fit since they can provide cut off values for assessing fit in structural equation modelling, there exists a degree of subjectivity and professional judgment in the selection of the appropriate model (Marsh, Byrne and Yeung, 1999) since reliance simply on "rules of thumb" can often lead to erroneous inferences (Marsh, Hau and Wen, 2004). For example, even though generally the model fit is regarded as non-acceptable when the chi-square statistic is significant, this is often not a fair judgment: The chi-square statistic depends on the sample size so when the sample size is large the statistic is almost always significant and the model is inappropriately rejected. In this respect, Marsh, Balla and Mc Donald (1988) stress the use of TLI in that, among other advantages, it has been found to be relatively independent of sample size. Bentler and Bonnett (1980) propose a value of .9 or higher for acceptable levels of fit. More recent thinking (Hu and Bentler, 1998) suggests that values above .95 are preferred. The same thresholds have been proposed for CFI. In relation to the RMSEA, values less than .05 have been considered indicative of a "close fit" while values up to .08 represent reasonable errors of approximation (Browne and Cudeck, 1993; Jöreskog and Sörborn, 1993).

Fit indices can be useful when comparing nested models; the focus is then on the observed increase or decrease in the corresponding fit index. A model is said to be nested in a less restrictive model when the set of parameters it involves is a subset of the parameters estimated in the less restrictive model. For the purposes of model comparison of the relative fit of nested models, the relative fit is more important than the absolute fit, as long as the best-fitting model is acceptable. When comparing each model with the immediately more restrictive one, support for the more constrained model can be justified when the decrease in TLI and CFI

is less than .01 (Cheung and Rensvold, 2001, 2002; Chen, 2007) and RMSEA increases by less than .015 (Chen, Curran, Bollen, Kirb and Paxton, 2008). It is important to emphasize that these are only rough guidelines (Marsh, Hau and Wen, 2004); a variety of different indices should be used for model comparison and factors such as theory, a priori predictions and common sense should all be taken into account before deciding on the best model to use.

3.3.8 A Summary of the Methodology Section

At this point, it should be clear to the reader that the data for my analysis were obtained by the PIPS monitoring project run by the CEM centre in Durham in an item level form and involved English primary school students. A sample was selected carefully from this database to reflect as accurately as possible the population of students and schools that participated in the year one and year four mathematics achievement assessments. The focus was on the youngest cohort of students in the database – namely those beginning primary school in 2004. For the sake of parsimony, the items for mathematics achievement were grouped together to form multiple indicators to control for measurement error. Multiple Imputation was used to fill in the missing values for mathematics achievement multiple indicators– five imputed datasets were obtained in this way. The obtained data provided information only at the level of the parcels, for mathematics achievement. However, the four different models from the 2x2 taxonomy, used for the purposes of my analysis, required the use of a range of a set of other variables. These varied depending on whether or not they made adjustments for measurement error – in which case multiple indicators should be incorporated – or not – in which case manifest scale scores were used. Moreover, the compositional models involved required that the school-level variables should be obtained by aggregating the student-level variables at the level of the school. The procedures followed for the derivations of the necessary variables have been described; I refer the reader to section A.3 in Appendix A for a more technical presentation on how these variables were used for the purposes of the analysis.

3.4 Results for Study 1a

The results of Study 1a are summarised in Table 3.3. In the next sections I expand on the main findings starting with the application of the conventional (doubly manifest) approach to the original data and then simulated data. I demonstrate how bias appeared in the estimated fixed effects and, crucially, in the estimated compositional effects simply as a result of adding random measurement error to the baseline scores. Importantly, I show how this bias was corrected for using appropriate models from the 2x2 taxonomy. I also address the impact of sampling error corrections on the estimates of compositional analysis. Finally I report the findings regarding the impact of multilevel measurement error on the estimates of the random effects of the compositional models that I incorporate in my analysis. Addressing relevant concerns in the educational effectiveness literature, I demonstrate how bias in the estimation of school effects in compositional value added models of educational effectiveness that involve compositional effects can be corrected for using latent models from the 2x2 taxonomy.

3.4.1 Results for Research Hypothesis 1a.1 and Research Question 1a.2: Applying the Doubly Manifest Approach to the Original Data

In this section I begin by summarizing the results of the conventional multilevel modelling approach, namely the doubly manifest approach to the original data. I remind the reader that the focus of enquiry was on the size and the direction of the compositional effect of school average mathematics achievement in year one (L2-MACH1) on individual achievement in year four (L1-MACH4) over and above the effect of individual-level mathematics achievement in Year one (L1-MACH1). The effect of L1-MACH1 on L1-MACH4 was predicted to be positive. The magnitude of the compositional effect was questioned for the analysis related to achievement (see also sections 2.3.6 and 2.3.7 in the literature review chapter that refers to the relevant on-going debates).

The Intra-class Correlation Coefficient (ICC) of the baseline measure (L1-MACH1) was reasonably high (ICC = .179), indicating that schools with varying levels of intake were involved in the sample (high achievement, well mixed and low achievement institutions). This was important for the present study in seeking to detect school compositional effects at the level of the school (Thrupp, Lauder and Robinson, 2002).

Individual-level achievement in year one was positively associated with subsequent individual achievement in year four (β_{within} = .685, se =.007; p<.001; ES = 1.263) consistent with my predictions (Research hypotheses 1a.1). A negative compositional effect of school average mathematics ability in year one on subsequent achievement in year four was found (β_{cont} = -.070, se = .032, ES = -.067). Although small, this negative effect contradicts the conventional wisdom that students achieve higher when going to higher achievement institutions (see also section 6.4.4 in the discussion chapter): it suggests that students who begin school with similar skills in mathematics, as these are measured by their test scores in year one, may make less progress if they attend higher attainment schools. In other words, the results suggest that having above average performing peers has, in fact, a detrimental effect on the individuals' acquisition of mathematics skills.

3.4.2 Results for Research Question 1a.3: Applying the Doubly Manifest Approach to the Simulated Data: Demonstrating Measurement Error Bias

In order to investigate in more detail the impact of measurement error on compositional effects estimates in my study I performed a small simulation study following Harker and Tymms (see section 1.7 in the introduction where I describe this study). When measurement error was deliberately added in the individual-level measures, a positive bias occurred in the estimated compositional effect (see Harker and Tymms, 2004; Marsh et al., 2010). Specifically, for data with reliability equal to .75, the compositional effect was found to be positive and non - significant β_{com} = .041 (se = .038, $ES_{\beta_{com}}$ = .036). When the reliability was further reduced to

.56, the compositional effect was estimated to be even more positive, so that it became highly significant β_{com} = .224 (se = .05, $ES_{\beta_{com}}$ = .158). These positive compositional effects were phantom effects and could only be attributed to the additional measurement error in year one achievement scores. It is clear that the larger the measurement error variability, the larger the bias in the estimated effects and for sufficiently low reliability the sign of the effect was reversed.

Measurement error in level 1 baseline scores resulted in a negative bias in the estimated within-group effect, that is, in the effect of individual-level prior achievement on subsequent achievement. Specifically, while the within-group effect was estimated to be equal to β_{within} = .685 (se =.007, ES = 1.263) in the analysis of the original data, it was found equal to β_{within} = .676 (se=.003, ES = 1.144) for data with reliability .75 and even smaller, equal to β_{within} = .627 (se=.009, ES = .976) for data with reliability equal to .56.

3.4.3 Results for Research Hypothesis 1a.4-Research Hypothesis 1a.6: Correcting for Bias due to Measurement Error Unreliability

This part of my analysis illustrates how the use of the latent manifest approach and the doubly latent approach can correct for measurement error bias in the estimation of the compositional effect.

The use of the latent manifest approach gave a more negative compositional effect than the doubly manifest approach when applied to the original data (β_{com} = -.132, se= .036, $ES_{\beta_{com}}$ = -.119). Thus, it effectively corrected for the positive bias in the estimate due to the occurrence of measurement error prevalent in the baseline scores. Crucially, the latent manifest approach corrected for measurement error bias even for data with substantially lower measurement error reliability than that of the original data. For the data with reliability .75, the

estimated compositional effect was β_{com} = -.174 (se = .047, $ES_{\beta_{com}}$ = -.13) and for the data with reliability .5, it was found β_{com} = -.177 (se = .071, $ES_{\beta_{com}}$ = -.117). In the same way, the doubly latent approach also eliminated bias in the estimated compositional effects. The estimates obtained were generally larger in magnitude than the effect of the latent manifest approach; the doubly latent approach makes further adjustments for sampling error (see also 3.4.4 on the results for Sampling Error Corrections).

The negative bias in the estimation of the within-group effect was also corrected for: this was dis-attenuated using the latent manifest and the doubly latent approaches (see Table 3.3).

3.4.4 Results for Research Question 1a.7: Sampling Error Corrections

Here we consider the role of sampling error in the compositional effect estimates. Adjustments for sampling error without further adjustments for measurement error (the manifest latent approach) were only considered in the analysis of the original data. These resulted in a more negative compositional effect β_{com} = -.079 (se= .037, ES = -.069) compared to that achieved using the doubly manifest approach (see previous section on the results of "Applying the doubly manifest approach to the original data"). Through the use of the doubly latent approach, adjustments for sampling error, in addition to measurement error, were considered. In all three analyses (the analysis with the original data and the two analysis with the simulated data) a more negative compositional effect estimate was obtained, compared to that obtained after only correcting for measurement error (see Table 3.3); this is in line with the findings of previous research (see section 2.2.3 of the literature review chapter entitled "Bias due to sampling Error"). Controlling for sampling error had generally no impact on the within group effects - sampling error only relates to aggregate level effects.

3.4.5 Results for Research Hypothesis 1a.8: The Impact of Measurement and Sampling Error Adjustments on Standard Errors

The approaches that I demonstrate can offer potential trade-offs in relation to bias and accuracy. This is depicted in Table 3.3 where the standard errors for the latent measurement approaches, and especially the doubly latent approach, are slightly larger than the standard errors associated with the estimates of the conventional doubly manifest approach. I note that for the same approach, standard errors associated with the within-group and the compositional effect estimates are also larger as measurement error reliability in the baseline becomes lower, correctly reflecting the uncertainty that is added in the level 1 measure. This is also illustrated in Table 3.3, comparing the standard errors for the estimates across the cells of each of the rows corresponding to the results for datasets with different level 1 measurement error reliabilities.

3.4.6 Results for Research Hypothesis 1a.9- Research Question 1a.11: The Impact of Measurement Error on Random Effects Estimates

The results underlying research questions 1a.9 and 1a.11 and research hypothesis 1a.10 are displayed in Table 3.4. The first column of the table displays the random effects estimates as obtained with the doubly manifest approach: the within-group variance, the between-group variance and the resulting school effect estimates are displayed in three distinct sub-columns. Moving across the cells of these sub-columns, the reader can observe the impact of adding random error to the baseline score on the corresponding random effects (Research Question 1a.9). Clearly, the residual variance at the within level was estimated to be larger for datasets with lower-level 1 reliability, reflecting the unreliability added in the level 1 scores. The level 2 variance remained relatively unaffected across the distinct reliability conditions; this is expected as no random error was added at level 2. School effect estimates were generally found to be slightly smaller for the datasets with lower reliability of the baseline scores, although the differences in the school effects estimates across the three datasets were not substantial.

The impact of measurement error adjustments on random effect estimates can be inferred by comparing the random effect estimates obtained with the doubly manifest approach (first column of Table 3.4) with those obtained with the latent manifest approach (third column of Table 3.4). For the latent manifest approach, the level 1 residual variance was estimated to be the same – and equal to .214 for all three datasets. Thus, bias due to measurement error on the within-group variance was corrected for applying the latent manifest approach. Correcting for this bias gave smaller level 1 residual variance as compared to that obtained with the doubly manifest approach that did not make adjustments for measurement error variance.

Finally, the impact of sampling error adjustments on random effects estimates can be seen by comparing the results of the doubly latent approach (fourth column of Table 3.4) with those for the latent manifest approach (third column of Table 3.4). Comparisons between the estimates of the manifest latent approach (second column of Table 3.4) with those of the doubly manifest approach (first column of Table 3.4) also allow the reader to make conclusions on the impact of sampling error adjustments on the estimated random effects. Generally controlling for sampling error did not lead to substantial changes in the estimates of level 1 and level 2 residual variance in the compositional models, and correspondingly, the estimated school effects were not altered after such adjustments. Still, with the doubly manifest approach, the within-group residual variance was estimated somewhat smaller as compared to the way it was estimated with the latent manifest approach. Additionally, the between-group variance being estimated a bit larger. Thus, school effects were estimated larger when simultaneous adjustments were made in the data for both measurement and sampling error (the doubly latent approach) as compared to when adjustments were made in the data solely for measurement error (the latent manifest approach).

3.4.7 A Summary of the Results for Study 1a

Study 1a, a large sample of English primary students in years one and four, revealed a positive relationship between individual achievement at year one and individual achievement at year four – this result is in agreement with conventional wisdom. Despite this, a weak negative but significant compositional effect of school average achievement at the end of year one, on individual mathematics achievement at the end of year four was detected. In a complementary study to that of the main analysis increasingly large random error was added to the baseline year one achievement. This led to a negative bias in the within-group effect - this was estimated smaller in magnitude – and a positive bias in the estimate of the compositional effect. For low enough measurement error reliability, the initially negative compositional effect became positive – an apparently phantom compositional effect. To compensate for measurement error bias in the compositional effect estimates, the latent manifest and the doubly latent model from the Marsh, Lüdtke et al. (2009) 2x2 taxonomy were used. These models eliminated bias both in the within-group effect and the compositional effect estimate, even for the simulated data – in spite of the fact that the latter data had substantially lower reliability than the original data. Lastly, according to the findings of Study 1a, measurement and sampling error adjustments in compositional models can also influence the estimation of random effects in compositional models. Although the impact of such adjustments on random effects is not strong, it is still of high relevance to the estimation of school effects as these are obtained from value added models of educational effectiveness (see section 6.4.5 in the discussion chapter).

Table 3.3:The impact of adjustments for measurement on the within effect and the compositional effect: Evidence from Study 1a.

Reliability	Approach followed							
	Doubly Manifest		Manifest Latent		Latent Manifest		Doubly Latent	
	$\beta_{within}\ (s.e.)$	$\beta_{com}\ (s.e.)$	$\beta_{within}\ (s.e.)$	$\beta_{com}\ (s.e.)$	$\beta_{within}\ (s.e.)$	$\beta_{com}\ (s.e.)$	$\beta_{within}\ (s.e.)$	$\beta_{com}\ (s.e.)$
	$ES_{\beta_{within}}$	$ES_{\beta_{com}}$	$ES_{\beta_{within}}$	$ES_{\beta_{com}}$	$ES_{\beta_{within}}$	$ES_{\beta_{com}}$	$ES_{\beta_{within}}$	$ES_{\beta_{com}}$
.875	.685 (.007)	-07 (.032)	.685 (.007)	-.079.(037)	.849 (.011)	-.132 (.036)	.848 (.011)	-.146 (.041)
	1.263 (.013)	*-.067 (.031)*	*1.287 (.013)*	*-.069 (.033)*	*1.424 (.015)*	*-.119 (.033)*	*1.451(.015)*	*-.121(.035)*
.75	.676 (.003)	.041 (.038)	.676 (.008)	.058 (.047)	1.035 (.017)	-174 (.047)	1.034 (.017)	-.198 (.054)
	1.144 (.013)	*.036 (.035)*	*1.165 (.014)*	*.021 (.017)*	*1.430 (.016)*	*-.145 (.04)*	*1.456 (.016)*	*-.150 (.041)*
.56	.627 (.009)	.224 (.05)	.627 (.009)	.309 (.065)	1.259 (.029)	-.177 (.071)	1.255 (.029)	-.220 (.077)
	.976 (.014)	*.158 (.036)*	*.992 (.014)*	*.187 (.04)*	*1.432 (.018)*	*-.117 (.048)*	*1.457(.019)*	*-.133 (.048)*

Note. The parameter β_{within} (standard error) denotes the within group effect of the outcome variable (mathematics achievement in year four) on the individual-level predictor (mathematics achievement in year one) while β_{com} (standard error) denotes the effect of school average achievement in Year one on mathematics achievement in Year four; ES denotes the Effect Size estimate for the corresponding effect.

Table 3.4: The impact of adjustments for measurement on the rando m effects estimates of value added models that make adjustments for prior achievement and average prior achievement

Reliability	Approach followed											
	Doubly Manifest			Manifest Latent			Latent Manifest			Doubly Latent		
	Within-group Variance	Between-group Variance	School effects Estimates	Within-group variance	Between-group variance	School effects Estimates	Within-group Variance	Between-group variance	School effects estimates	Within-group variance	Between-group Variance	School effects estimates
.875	.319	.072	18.4%	.319	.072	18.4%	.214	.07	24.6%	.213	.072	25.3%
.75	.376	.073	16.3%	.376	.073	16.3%	.214	.069	24.4%	.210	.073	25.8%
.56	.446	.077	14.7%	.446	.075	14.4%	.214	.066	23.6%	.208	.074	26.2%

Phantom Compositional Effects: Evidence from Cyprus (Study 1b)

In Study 1b, I use data on mathematical achievement from year four students in Cyprus obtained from the project "Promoting Quality in Education" (see www.ucy.ac.cy/esf). In this study, the substantive focus is to investigate how the school average achievement in mathematics can affect the student's progress in mathematics at the fourth year of their primary education. Methodologically, I am concerned with the extent to which measurement error is a source of a "phantom" compositional effect in this data. The way in which sampling error alters compositional effects estimates is also of interest. Appropriate models from the 2x2 taxonomy are used throughout the analysis. In this respect, Study 1b enables the evaluation of the generalizability of the results in a very different educational setting. The main focus is not on the sizes of the effects per se but more on the usefulness of the models and their application in different contexts.

3.5 Research Questions and Research Hypotheses for Study 1b

Study 1b evolved around two main research questions:

Research Question 1b.1: What is the size of the effect of the individual mathematics achievement and school average achievement in the beginning of year four on subsequent achievement at year four?

Research Question 1b.2: How do the estimates of the individual and the compositional effects change after adjustments for measurement error (the latent manifest approach) or measurement and sampling error simultaneously (the doubly latent approach)?

The focus here was on the extent to which measurement error was a source of phantom compositional effects with these data. Predictions could be made in a similar manner as for Study 1a, in which I also investigate compositional effects, using English data (see sections 3.2.3 on "Correcting for Bias due to Measurement Error Unreliability" and 3.2.4 on "Correcting

for Sampling Error" under the research questions section for Study 1a). The expectation was that adjustments for measurement error would correct for the positive bias in the estimated compositional effect, while simultaneous adjustments for measurement and sampling error (application of the doubly latent approach) would lead to stronger compositional effects estimates, as opposed to when adjustments are made only for measurement error (the latent manifest approach).

3.6 Methodology for Study 1b

3.6.1 Measures and Data Samples

The tests measuring students' mathematics achievement, incorporated for the purposes of the present investigation, were administered at the beginning and end of school year 2010-2011 to all grade four students from a sample of primary schools in Cyprus. My analysis was based on data from 1,694 students, who participated in both measurement sessions. There were a total of 59 schools in my sample.

For the construction of the tests, permission was obtained from IEA to use the released items of TIMSS 2007. The properties of each item and the relation with the curricula of years three and four in Cyprus were taken into account for developing four parallel types of test in each subject (Panayiotou et al., 2013). Thus, the data were available both at the item level and as total scores.

3.6.1.1 The "Dynamics of Educational Effectiveness Research" ("ESF") Project

The data were originally collected as part of a larger project, namely the "Dynamics of Educational Effectiveness Research" ("ESF") project (www.ucy.ac.cy/esf). Six European countries participated in this project including Cyprus. One of its main objectives was to investigate differences between schools in the added value of primary education for different outcomes of schooling. It also sought to identify factors operating at different levels, for

example at the level of the student or the school which could explain these differences between schools and their added value. To be precise, it tested the validity of the dynamic model of educational effectiveness (Creemers and Kyriakides, 2008) by searching for the effects of factors included in the model upon student achievement at the end year four.

3.6.1.2 Design

Test booklets, as administered in the TIMSS 2007 assessment cycle were used. Because not all test booklets from the TIMSS 2007 cycle were released for public use, only the booklets one to four were administered. The booklets, items and codes for the first time point were exactly the same as the second point of measurement; the booklets were just rotated as displayed in Table 3.5.

Table 3.5: The booklet rotation design used in the administration of the tests for the "Dynamics of Educational Effectiveness Research Project"

Beginning of the year entering grade four	**End of grade four**
Test booklet 1	Test booklet 3
Test booklet 2	Test booklet 4
Test booklet 3	Test booklet 1
Test booklet 4	Test booklet 2

3.6.1.3 The use of the Belgian SIBO items

The books used for the "Dynamics of Educational Effectiveness Research" project differ from the TIMSS test booklets in two aspects: First, it was decided not to administer a certain set of test items because additional supplementary material would have been necessary for their administration. Secondly, because of concerns that some TIMSS-items might be too difficult for early grade four students, the TIMSS test blocks were supplemented with twelve items originating from the Belgian "*Schoolloopbanen in het BasisOnderwijs"* (SIBO).

3.6.2 Missing Data

The main problem in the statistical analysis was the treatment of item non-response. In the statistical analysis, several different types of non-response could be identified (see Vennemann and Wendt, 2012):

Omitted Items: An item was characterised as "omitted" when a student who should have answered the item did not, when an item was left blank or when two or more response options were checked for a multiple choice item.

Non-reached items: An item was considered as not reached when the item itself and the item immediately preceding it were not answered, while there were no other items completed in the remainder of that part of the booklet.

Non-administered items: These are items that were not administered, either by design, belonging to a booklet other than the one originally used by a student or unintentionally when an item was misprinted or otherwise unavailable for a student to answer.

Ambiguous-items: For such items were it was not clear what answer was given by the student. For example, for a multiple choice question the circle could be somewhere between the two choices, making it difficult for the marker to decide which answer was chosen by the student.

For the purposes of my analysis, I replaced omitted and non-reached items with the probability of answering the item correctly. For a multiple choice item, this probability depends on the number of choices available to the student. For the TIMSS multiple choice items, there were four options to choose from and therefore the probability of giving the correct response was .25. For the SIBO multiple choice items there were five choices to choose from, thus .2 was the value used to replace omitted and non-reached items. Lastly, for open ended questions, the probability of answering the item correctly was taken to be equal to zero. Not administered and ambiguous items were treated as system missing. Thus, it was assumed that no answer was

given by the student to the specific item. This was important in the calculation of multiple indicators measuring the latent mathematics achievement for each student (see section 3.6.3 on the variables used for the purposes of Study 1b).

3.6.3 Variables for Study 1b

The data were originally organised in three different data files: One with the scale scores for both measurement sessions, another with item-level data form measurement occasion one and another with item-level data for the second measurement session. The three data-files were merged based on the student identification number. In this section I describe how I formed the scale scores and the multiple indicators for students' mathematics achievement based on the available data.

3.6.3.1 Scale scores

The overall scale scores for each student were available in the database, collected as part of the "ESF" project (see section the "Dynamics of Educational Effectiveness Research" project). Specifically, I used the variables WLE_mn_L_I and WLE_mn_L_II, as recommended by the User Guide (Vennemann and Wendt, 2012): the intention was to investigate differences in the mathematics achievement of the students between the first and the second measurement occasion. Further, I intended to interpret only the results of my analysis at the national-level, relative to the Cypriot sample.

3.6.3.2 Multiple Indicators

Multiple indicators for each student were constructed by taking the average score across the items corresponding to the same content area. In this way, three different indicators were obtained for each student: One for Numeric ability, one for Geometry and one for Data Display. Since the multiple indicators were not already available in the database, I followed certain procedures to obtain them; in what follows I describe these procedures in more detail.

3.6.3.3 Classification of the items according to their content domain

All the items were already classified according to content domain (Number, Geometric Shapes and Measures, Data Display) on the basis of the TIMSS 2007 Framework (Foy and Olson, 2009). I based the construction of multiple indicators for the students' mathematics achievement on this classification (Panayiotou et al., 2013).

The items were distributed in a homogeneous way across the booklets; nevertheless, there existed a substantively larger number of items classified as "numeric" compared to the number of items classified as "geometry" or "data display" (see also Table 3.6).

Table 3.6: The distribution of the items across booklets and across content areas

	Items classified as coming from…	Items number in Booklet 1	Items number in Booklet 2	Items number in Booklet 3	Items number in Booklet 4	Items number (total)
Origin	**SIBO**	4	2	3	3	**12**
	TIMSS 2007	21	19	23	25	**55**
	Total	**25**	**21**	**26**	**28**	**68**
Content Area	**Numeric**	16	14	14	14	**41**
	Geometry	5	4	8	8	**16**
	Data Display	4	3	4	6	**11**
	Total	**25**	**21**	**26**	**28**	**68**

3.6.3.4 Standardization of each of the multiple indicators for mathematics achievement

The scores for each of these three domains were based only on the items in the booklet that the student completed in each measurement occasion. This means that they were not computed using the same set of items for each student. This could be problematic in that the items might

not be equivalent across the four booklets. For example, the items could be more difficult in one booklet compared to another.

To compensate for this, the scores for the numeric, geometry and data display ability of the students were standardized in the following way: A separate mean score and standard deviation estimate was obtained for each booklet, based on all those students who completed the specific booklet in either the first or the second measurement session. These estimates were used to standardize the scores of the students who completed the corresponding booklet at either of the two time points. For example, if a student completed booklet one in measurement occasion one, then the standardized score for this student was obtained by subtracting the mean score (over all those student who completed booklet one in either the first or the second measurement session) and dividing by the standard deviation (of the same group of students as the one based on which the mean was calculated) corresponding to booklet one. The same mean and standard deviation was used to obtain the standardized scores for a student who completed booklet one in measurement sitting two. In this way, the scores across the four booklets were placed in the same metric.

3.6.3.5 The final dataset

In the final dataset the following information was available for each student: (i) the school identification number, (ii) the overall scale score for each measurement session (see section 3.3.4), (iii) the scores of the students at the three cognitive domains (Numeric, Geometry and Data Display) both at the beginning and at the end of grade four, (iv) the school-level variables, obtained by aggregating the individual-level variables across all students within the same school. School aggregates were obtained both for the overall scores and the multiple indicators.

It should be noted that all the level 1 variables were standardized; the level 2 variables were based on these standardized scores but not re-standardized themselves, just like the analysis with the English Performance Indicators in Primary School (PIPS) project (see section "Standardisation" under 3.3.4 for Study 1a).

3.7 Results for Study 1b

3.7.1 Results for Research Hybothesis 1b.1 and Research Hypothesis 1b.2

The application of the conventional compositional analysis model (doubly manifest approach) revealed a positive and significant compositional effect (β_{com} = .186, se = .079, $ES_{\beta_{com}}$ = .137). However, this apparently positive compositional effect disappeared after adjustments for measurement error. Specifically, when the latent manifest approach was used, the effect was found to be smaller and became non-significant (β_{com} = -.023, se = .118, $ES_{\beta_{com}}$ = -.018). The compositional effect remained non-significant even after additional adjustments for sampling error through the use of the doubly latent approach (see Table 3.7). It should be noted that the intra class correlation (ICC) of the baseline score– that shows the unadjusted between school differences in the students' ability – was only equal to .08 for the Cypriot data indicating moderate differences across schools in their intake. This was much lower than the corresponding ICC for the English data that was estimated to be equal to .179 (see section in the methodology for Study 1a in which this results is reported). Therefore, it would be more difficult to detect significant compositional effects using the data from Cyprus; in order to detect school compositional effects the requirement is to have schools at the extremes of the intake: only in this way will we be able to gain access to the schools at the extremes of the prior achievement distribution (see, for example, Thrupp et al., 2002).

Moreover, in accordance with the findings of the simulation study, correcting for measurement error (latent manifest approach) also led to larger estimates of the within-group

effect as compared with the conventional approach. Further adjustment for sampling error (the doubly latent approach) did not alter substantively the within-group effect (see Table 3.7).

Lastly, the bias accuracy trade-off is also prevalent here: even though measurement error bias in estimation is eliminated using the partial and full correction approaches, the standard error associated with the estimates is slightly larger (see Table 3.7).

3.7.2 A Summary of the Results for Study 1b

Study 1b, a large study of primary students in year four, revealed a small positive compositional effect of school average achievement at the beginning of year four on subsequent individual achievement at the end of year four. After adjustments for measurement error, through the use of the latent manifest approach from the 2x2 taxonomy, this positive effect – apparently phantom in nature – became negative and non-significant. Further adjustments for sampling error led to a more negative but still non-significant compositional effect.

Table 3.7: Application of the 2x2 taxonomy of models to the data from Cyprus

	β_{within} $(s.e.)$ $ES_{\beta_{within}}$ $(s.e.)$	β_{com} $(s.e.)$ $ES_{\beta_{com}}$ $(s.e.)$
Doubly Manifest	.654 (.019) *1.224 (.034)*	186. (.079) *.137 (.058)*
Manifest Latent	.654 (.019) *1.260 (.031)*	.293 (.118) *.157 (.061)*
Latent Manifest	.924 (.096) *1.401 (.094)*	-.023 (.143) *-.018 (.109)*
Doubly Latent	.933 (.099) *1.467 (.095)*	-.068 (.285) *-.034 (.141)*

Note. The parameter β_{within} denotes the within group effect of the outcome variable (mathematics self-concept in Year four) on the individual-level predictor (mathematics achievement in Year one) while β_{cont} denotes the effect of school average achievement in Year one on mathematics self-concept in Year four. The symbol $(s.e.)$ denotes the standard error of the parameter estimate and of the corresponding effect size; ES denotes the Effect Size estimate for the corresponding effect.

Chapter 4: The Big Fish Little Pond Effect: Evidence from Early English Primary School Data (Study 2)

4.1 Introduction

For the purposes of my analysis, in the present study (Study 2 of my thesis) I used the same dataset as in Study 1a, namely primary year one and four year longitudinal data from the Performance Indicators at Primary School Project (PIPS; see section 3.3.1). As in the first study of my thesis (Study 1), in Study 2 I was also concerned with compositional effects of school average mathematics achievement on students' educational outcomes – however, the focus was on an affective educational outcome, namely on self-concept.

Using a psychological perspective, Marsh (1987) found that school average achievement negatively influenced academic self-concept, despite the positive relationship between individual achievement and self-concept. This finding, which describes the Big-Fish-Little-Pond-Effect (BFLPE), has been verified in numerous studies since then (see Marsh, Seaton et al., 2008; Seaton and Marsh, 2010) and has been well-established in educational psychology.

Study 2 has two substantive purposes. The first (Study 2a) is to extend BFLPE studies to younger children throughout their first four years of schooling. The second (Study 2b) is to evaluate whether the big-fish-little pond effect might explain, in part, the negative effect of school-average achievement on subsequent mathematics achievement that was detected in Study 1a (see section 3.4 in which the results of Study 1a are reported). Where this contributes to existing literature is that both components of Study 2 confirm theoretical models underlying the BFLPE paradigm using data from children in early primary school years. Methodologically, the original contribution to knowledge is the use of multilevel structural equation models to correct for measurement error (and/or sampling error) bias in the estimation of BFLPEs.

4.1.1 Study 2a: The BFLPE in Early Primary School Years

In the first component of Study 2 (Study 2a) I investigate the generalizability of the BFLPE across samples of students of different chronological ages. Specifically, I test (i) the compositional effect of school average achievement at the end of year one (L2-MACH1) on students' self-concept at the end of year one (L1-MSC1), (ii) the compositional effect of school average achievement at the end of year one (L2-MACH1) on students' self-concept at the end of year four (L1-MSC4).

The BFLPE for students as young as those considered in my sample has been considered by Tymms (2001) who found negative effects of school average achievement on subsequent self-concept in mathematics, reading and school. Despite the fact that these effects were significant, they were rather weak (see "Generalizability of the BFLPE across students of different age" in section 2.1.5 of my literature review). Given the large sample of students involved in this study (21 000 pupils taught in 628 schools), the doubly latent model (see section in the literature review where I describe the doubly latent and the latent manifest models) could arguably be used with these data without serious concerns about convergence. Additionally, given the reported reliability of the scale measuring attitudes (e.g. .6 for the scale measuring attitudes towards mathematics) the expectation would then be that the application of the doubly latent models would lead to substantially larger BFLPEs (see "Measurement error in explanatory variables at level 1" in section 2.2.3). This hypothesis is addressed in my thesis (see Study 2a) using data from the same monitoring project as that in Tymms' study (PIPS), namely for students ranging from five to seven years of age.

4.1.2 Study 2b: Academic Self-Concept as Mediator of the Negative Compositional Effects of School Average Achievement on Students' Subsequent Self-Concept and Students' Mathematics Achievement

In Study 2b I initially investigate the stability of the BFLPE through years one and year four of primary education. Specifically, the focus is on the effect of school average achievement at the end of year one (L1-MACH1) an academic self-concept at the end of year four (L1-MSC4), after controlling for academic self-concept at the end of year one (L1-MSC1).

Then, bridging the findings of Study 1a (a negative compositional effect of L2-MACH1 on L1-MACH4) with Study 2a (a negative compositional effect of L2-MACH1 on L1-MSC4), I examine whether the negative compositional effect of L2-MACH1 on L1-MACH4 (Study 1a) can justifiably be attributed to the prevalence of the big-fish-little-pond-effect evident with these data (Study 2a). Specifically, I examine the extent to which academic self-concept in year four (L1-MSC4) mediates the negative compositional effect of school average achievement at the end of year one (L2-MACH1) on students' achievement at the end of year four (L1-MACH4), as this was found in Study 1a.

The critical outcome in BFLPE studies is academic self-concept - in contrast with research in the field of educational effectiveness, which is primarily interested in compositional effects of school average achievement on cognitive outcomes of schooling and subsequent educational achievement (see substantive focus of Study 1a and Study 1b), Nevertheless, the implications of this phenomenon for academic achievement and performance have repeatedly been questioned (see for instance Marsh, Seaton et al., 2008). Indeed, sociological models of school composition (see Marsh, 1991) typically assume that psychological processes mediate the effects of school-average achievement on academic outcomes (e.g. Alwin and Otto, 1977). Marsh (1991), applying the theoretical model of Bandura's theory of social cognition (Bandura, 1986) in the educational context, posits that school composition may influence academic self-concept directly formed during early school years. These initial effects then mediate the

influence of school composition on subsequent academic outcomes during later school years and eventually on educational achievement. Marsh tested and verified his theory using high school longitudinal data of year ten, twelve and two years after graduation from high school. Following this study and in accordance with Bandura's theory, I also investigate the extent to which academic self-concept mediates school compositional effects of school average achievement on subsequent academic achievement (Study 2b) – but this time with early primary school data.

4.2 Research Questions and Research Hypotheses Underlying Study 2a: The BFLPE in Early Primary Years

4.2.1 The Compositional Effect of Year One Mathematics Achievement on (i) Year One and (ii) Year Four Mathematics Self-Concept Using the Doubly Manifest Approach

The BFLPE hypothesis, the focus of this study, predicts that the compositional effect of school average achievement on subsequent self-concept in the corresponding academic domain is negative, while individual prior achievement is positively correlated with subsequent self-concept (see section 2.1.3 in the literature review in which I refer to the BFLPE hypothesis). Previous research on the BFLPE has been consistent in suggesting that the BFLPE is generalizable across different groups, academic domains and educational systems (Marsh, 1991; Marsh and Craven, 2002; Marsh and O' Mara, 2010; Marsh, Trautwein, Lüdtke, Baumert and Köller, 2007; see review by Marsh, Seaton, et al., 2008). Among the main issues that BFLPE research seeks to establish is the generalizability of the phenomenon for students of different ages and its stability across time for students who remain in the same school setting for several years. Evidence on these two characteristics of the BFLPE comes from data from secondary school and late high-school grades (see section 2.1.5 in the literature review on the generalizability and on the stability of the BFLPE). Importantly, however, the BFLPE has not

previously been evaluated for children as young as five to seven years old (but see Tymms, 2001). This research is, in part, a response to the need to investigate the extent to which the BFLPE can be generalized for younger students in their early primary school years. My hypothesis was that this phenomenon would be verified even for this age range.

In two separate analyses, I investigate (i) the effect of school average achievement in year one (L2-MACH1) on individual mathematics self-concept in year four (L1-MSC1) over and above the effect of individual mathematics achievement in year one (L1-MACH1), (ii) the effect of school average achievement in year one (L2-MACH1) on individual mathematics self-concept in year four (L1-MSC4) over and above the effect of individual mathematics achievement in year one (L1-MACH1).

Research Hypothesis 2a.1: The effect of individual mathematics achievement in year one (L1-MACH1) on both mathematics self-concept in year one (L1-MSC1) and mathematics self-concept in year four (L1-MSC4) is positive. The effect of school average mathematic achievement in year one (L2-MACH1) on both outcomes is negative, confirming the BFLPE hypothesis with year one and year four data.

4.2.2 Using Partial and Full Correction Approaches for the Investigation of the BFLPE

Perhaps one of the reasons why educational researchers have been hesitant to address the BFLPE with early primary school data has been the fact that younger children may have difficulty in completing self-report questionnaires and, therefore, data on self-concept in these year groups can be especially unreliable. The use of doubly latent models with research of this nature – dealing with data from rather young students – can therefore be considered a necessity: through the use of this approach taking into account measurement error in the data becomes possible.

In this section I form the research question and research hypotheses on how the within group effect, the compositional effect and associated standard errors change when different models from the 2x2 taxonomy are used (and thus, different assumptions for the prevalence of measurement and sampling error are made in the data). I refer the reader to the relevant sections in the literature review where I discuss the impact of measurement and sampling error on compositional analysis estimates (section 2.2.3) and to the sections in Study 1a where I form analogous research questions and research hypotheses for an explanation of the hypotheses made here (section 3.2. and, sepcifically, section 3.2.3, section 3.2.4 and section 2.3.5).

The set of Research Hypothesis 2a.2 up to Research Hypothesis 2a.4 refers to both the analysis with L1-MSC1 as the outcome measured (verification of the BFLPE with year one data) and to that in which L1-MSC4 is used as the criterion (verification of the BFLPE in year four).

Research Hypothesis 2a.2: Adjusting for multilevel measurement error through the models of the 2x2 taxonomy lead to a stronger (more negative) compositional effect of achievement on self-concept. Meanwhile the within group effect becomes more positive.

Research Hypothesis 2a.3: Furthermore, adjustments for sampling error lead to even stronger compositional effects; that is, the negative compositional effects become more negative. However, sampling error does not affect the estimation of the within group effect.

Research Hypothesis 2a.4: The partial and full correction models of the 2x2 taxonomy offer potential trade-offs in relation to bias and accuracy: While measurement error bias in estimation is eliminated, the estimation of the effects is less powerful in that the standard errors are larger.

4.3 Research Questions and Research Hypotheses Underlying Study 2b: Academic Self-Concept as Mediator of the Negative Compositional Effects of School Average Achievement on Students' Subsequent Self-Concept and Students' Mathematics Achievement

4.3.1 The Growth of BFLPEs over the First Four Years of Primary Schooling

Extending the compositional model that I used to verify BFLPE with year one and year four data, I investigated the stability of the BFLPE through years one to four of primary school; stability refers mainly to the relative size of the BFLPE across the distinct phases of schooling. Empirical evidence suggests that students who remain in the same school setting for several years experience more negative big-fish-little-pond-effects with time (see section 2.1.5 in the literature review on the BFLPE).

Research Hypothesis 2b.1: The compositional effect of L2-MACH1 on L1-MSC4 is stronger (more negative) compared to the compositional effect of L2-MACH1 on L1-MSC1.

Research Hypothesis 2b.2 refers to a mediation model in which the total effect of L2-MACH1 is conceptualized as the sum of potential direct effects of L2-MACH1 on L1-MSC4 and indirect effects of L2-MACH1 on L1-MSC4 through L1-MSC1.

Research Hypothesis 2b.2: There are additional BFLPEs the longer the student remains in the school so that, in testing a model with L1-MSC1 mediating the effects of L1-MACH1 and L2-MACH1 on L1-MSC4, the direct effect of L2-ACH1 on L1-MSC4 is negative.

If no direct effects are found, this would mean that the compositional effect of L2-MACH1 on L1-MSC4 as compared to the effect of L2-MACH1 on L1-MSC1 becomes neither larger nor smaller. A negative direct effect –one of the same direction of the direct compositional effect - means that the effect becomes more negative over time while a positive indirect effect – one of the opposite direction of the direct composition - would mean that the compositional effect decreases over time in magnitude (in the case of big-fish-little-pond-effects it becomes more negative).

4.3.2 Extending the BFLPE: Is the Negative Compositional Effect of School Average Achievement on Subsequent Achievement (Study 1a) Mediated by the BFLPE?

In this final analysis of Study 2, bringing together the findings of Study 2a and the findings of Study 1a, I investigate whether the negative relationship between school average mathematics achievement in year one and mathematics achievement in year four is mediated by mathematics self-concept in year four.

The major focus of BFLPE studies have been on ASC as the main outcome variable. Nevertheless, a number of studies suggest that some of the effects of school average ability on other constructs are at least partially mediated by academic self-concept (Seaton, Marsh and Craven, 2009; Trautwein et al., 2006). Particularly in relation to the effects of school average achievement on subsequent achievement, Marsh (1991) found that the negative effects of school average achievement on standardized test scores were substantially reduced after controlling for the negative effects of school average ability on academic self-concept. Similarly, Marsh and O' Mara (2010), using the Youth in Transition data found that academic self-concept mediated the negative effects of school average achievement on students' subsequent school grades and grade point average. Nagengast and Marsh (2012) reported that the negative effects of school-average achievement on academic aspirations are substantially mediated by academic self-concept.

Based on this research, an immediate hypothesis naturally follows regarding the negative compositional effects of school average achievement on subsequent student mathematics achievement detected in Study 1a.

Research Hypothesis 2b.3: The indirect effect of L2-MACH1 on L1-MSC4 is negative and significant: L1-MSC4 mediates the negative effect of L2-MACH1 on L1-MACH4 as detected in Study 1a.

4.3.3 Summary of the Research Questions and Research Hypotheses for Study 2

Study 2 investigates the compositional effect of year one mathematics achievement on year one mathematics self-concept and year four mathematics self-concept (Study 2a). In this way, it seeks to verify the prevalence of big-fish-little-pond-effects with the PIPS mathematics achievement dataset that was used in Study 1a. The hypothesis is that the BFLPE can be verified for students in this age range. Methodologically, the focus is on the use of the partial and full correction models of the 2x2 taxonomy (see section 2.2.5 in the literature review) with appropriate adjustments in the modelling – where necessary -- to address the research question and research hypotheses posed. In a separate analysis (Study 2b), the growth (stability) of the BFLPE over time, throughout the first four years of schooling, is also addressed. Crucially, Study 2b, extending the BFLPE and bridging Study 1a and Study 2a, tests the extent to which academic self-concept in year four, mediates, at least partially, the negative compositional effect of school average achievement in year one on an individual's achievement in year four, as this was detected in Study 1a.

4.4 Methodology

The methodology followed in Study 2 in testing compositional effects of school average achievement on individual self-concept (big-fish-little-pond-effects) was similar to that described in Study 1 (see section 3.3 where I describe the measures and procedures and the application of the four models of the 2x2 taxonomy in Study 1) in investigating compositional effects of school average achievement on subsequent achievement. This involved treating missing data in the appropriate manner, deriving the variables required for the application of the models of the 2x2 taxonomy based on the item-level data existent in the dataset available and, finally, running the appropriate Mplus codes to obtain the intended compositional effect estimates of interest. I expand on relevant issues in the sub-paragraphs of the present section.

4.4.1 Measures and Data Samples

The achievement measures incorporated for Study 2 came from the same English database as the first study presented in my thesis (see section 3.3.2 on the measures and data samples used in Study 1a); the data sample was that used for the purposes of Study 1a. For the purposes of Study 2, I additionally used mathematics self-concept measures for year one and year four students available in the database.

The self-concept measures provided by the PIPS tests consisted of five items. These were designed to assess the attitudes of the pupils towards mathematics. Each item had a choice of four options to choose from. The statements could be characterized as hybrids of attitude and self-concept measures but, for the purposes of the present study, all the items were treated as self-concept measures. The reliability as estimated by the Cronbach's alpha for year one self-concept measures was estimated to be equal to .614 and for year four self-concept measures equal to .716.

4.4.2 Missing Data

4.4.2.1 Defining unit non-response

I remind the reader that in the data sample incorporated both for Study 1a, students who did not participate in year one or year four PIPS examination were treated as missing units (see section 3.3.3 on the treatment of missing data in Study 1a). In contrast with the items for the mathematics achievement tests in PIPS (see "Defining unit non-response" under section 4.4.2 in Study 1a) there were no serious problems related with item non-response for mathematics self-concept measures, inasmuch as most students who participated in either year one or year four assessment completed all of the relevant items (see Table 4.1). So no further cases were treated as missing units because of inadequacies in the self-concept data.

4.4.2.2 The use of Multiple Imputation to treat missing data

Missing data in the students' self-concept measures (both unit non-response and item non-response) were imputed. The self-concept measures for year one and year four were included in the same imputation model as the mathematics achievement measures (see section "The use of Multiple Imputation to treat missing data" in section 3.3.3 of Study 1a). Having obtained the imputed data from the analyses in Study 1a, it was therefore unnecessary to proceed with further adjustments for missing data for the purposes of Study 2: I used the data as these were also used for the analyses of Study 1a.

Table 4.1: Number of students who attempted all, some or none of the items related to the self-concept measures employed in the study

Year group	Num-ber of items	valid cases	No items attempted	One item completed	Two items completed	Three items completed	Four items completed	Five items completed
Year one	Five	16974	893 (5.3%)	21 (.1%)	35 (.2%)	126 (.7%)	528 (3.1%)	15371 (90.6%)
Year four	Five	17352	174 (1%)	2 (0%)	3 (0%)	4 (0%)	95 (5%)	17074 (98.4%)

Note. Five items were used in my datasets to measure mathematics self-concept. The proportion of students answering different number of items for Year one and Year four data is displayed in this table.

4.4.3 Variables

Following the imputation procedure, five distinct imputed datasets were produced (see section "Maximum Number of iterations and number of produced datasets" in section 3.2.3 for Study1a). Just as mathematics achievement measures (see section 3.3.4 on "Variables" for Study 1a), the data on mathematics self-concept measures were in the form of multiple indicators – items measuring student-level mathematics self-concept. These were manipulated accordingly to derive the student and school-level variables necessary for the models incorporated (see Table 2.1 displaying the four models of the 2x2 taxonomy).

4.4.3.1 Student-level variables

For the doubly manifest and the latent manifest approach the scale score for mathematics self-concept was estimated as the average of the imputed self-concept items. The items themselves were used as multiple indicators for the latent manifest and the doubly latent approach.

4.4.3.2 School-level variables

The school-level variables were obtained by taking the school-level average of the student-level scale variables (for the doubly manifest and the manifest latent approach) or items (multiple indicators; for the latent manifest and the doubly latent approach) assuming manifest or latent aggregation, depending on whether or not adjustments for sampling error were made (see "School-level variables" under the section 3.3.4 for Study 1a).

4.4.4 Statistical Analysis

4.4.4.1 The use of the 2x2 taxonomy to investigate big-fish-little-pond-effects

Compositional effects of school average mathematics achievement on students' self-concept in relation to mathematics (Research Hypothesis 2a.1-Research Hypothesis 2a.4) were all addressed using the four models of the 2x2 taxonomy. The extent to which statistical inferences altered when different assumptions were made for measurement error with or without further adjustments for sampling error was investigated.

4.4.4.2 The treatment of the multilevel structure in the data as a nuisance[3] factor in investigating 2-1-1 mediation

In investigating mediation in Study 2b (see Research Hypothesis 2b.1, Research Hypothesis 2b.2), I was interested in the effect of an independent level 2 variable (school average mathematics achievement at the end of year one) on a level 1 mediator (mathematics self-concept at year one) which, in turn affected a dependent level 1 variable (individual mathematics self-concept in year four). In the same way, in Study 2b (see Research

[3] The use of the term nuisance here is technical, and refers to the fact that the statistical procedures followed adjust standard errors accordingly without decomposing the variance of the variables involved into "within" and "between" components.

Hypothesis 2b.3) I addressed the effect of school average achievement at the end of year one on mathematics self-concept at year one which, in turn, affected mathematics achievement at the end of year four. Both these scenarios are examples of a 2-1-1 design in educational psychology.

Generally, in 2-1-1 research designs the researcher is not interested in assessing relationships at the between level separately from relationships at the within level. The use of multilevel modelling is thus unnecessary in this situation, if meaningful at all (see Lüdtke, Robitzch, Thoemmes and Trautwein, 2013). For instance, and, with respect to my analysis, aggregated self-concept is measuring something different rather than individual self-concept and it is not clear what theoretical construct aggregated self-concept should be measuring. Therefore, following recommendations by Lüdtke et al. (2013), I treat the presence of hierarchical data as a nuisance factor in assessing indirect effects in my analysis.

The treatment of clustered data as a nuisance factor is an alternative procedure to multilevel modelling to accommodate for the impact of the presence of correlated observations on the standard errors of the estimates obtained by relevant analyses. It involves applying a single-level mediation model (OLS approach) while adjusting the standard errors to account for the fact that they can be underestimated leading to liberal significance testing (Kish, 1965). In a simulation that is based on real data, Lüdtke et al. (in, press) demonstrate that, when the unit of inference is the individual, such as in 2-1-1 designs, the estimates of the indirect effects obtained by the OLS approach with clustered standard errors are more robust in terms of model specification than those of alternatives (see, for instance, Preacher, Zyphur and Zhang, 2010) that separate the effects of variables into within- and between- group components.

4.4.4.3 The use of the "MODEL CONSTRAINT" command in Mplus to assess mediation

In assessing mediation for the purposes of my analyses, I examined total effects, direct effect and indirect effects.

With respect to research hypothesis 2b.1, I assessed the effect of L2-MACH1 on L1-MSC4, after adjustments for L1-MACH1 and L1-MSC1 (direct effect) and the indirect effect of L2-MACH1 on L1-MSC4 via L1-MSC1. A statistically significant and negative direct effect of L2-MACH1 on L1-MSC4 would imply that the negative effects of school average mathematics achievement at the end of year on individual self-concept persisted up until the fourth year of primary education, even after adjusting for big-fish-little-pond-effect at year one (negative effect of L2-MACH1 on L1-MSC1).

With respect to research hypothesis 2b.2, the focus was on the effects of L1-MACH1 and L2-MACH1 on L1-MASC4 after adjustments were made for L1-MASC1 in the models (direct effects) as well as on the indirect effect of L1-MACH1, and, most importantly, L2-MACH1 on L1-MACH4 via L1-MSC4.

A statistically significant indirect effect of L2-MACH1 on L1-MACH4 would suggest that the negative compositional effect of L2-MACH1 on L1-MACH4 was at least partially mediated by big-fish-little-pond-effects on self-concept. The total effects were conceptualized as the sum of the direct (unmediated) and indirect effect (mediated) of the corresponding variable on L1-MACH4.

The mediation model was specified in Mplus using the complex modelling procedure (Muthén and Muthén, 2012; see Appendix D.1 for the syntax). This option could effectively handle the non-independence of the scores from students coming from the same school, providing standard errors estimates corrected for the clustering effect of students nested within schools. Structural Equation Models were incorporated, using multiple indicators for the latent constructs involved in the analysis. For L1-MACH1 and L1-MACH4 these were the parcels formed based on the mathematics achievement items of the corresponding test (see section 3.3.4 of Study 1a). For L1-MASC4 the self-concept items for year four were incorporated (see section 4.4.1). For L2-MACH1 manifest aggregation was followed; the level 2 multiple indicators were formed by averaging the items at the school-level.

Direct effects were estimated as the standardized beta weights of the corresponding variable (L1-MACH1 or L2-MACH1) when all variables were included in the model (see also Marsh and O'Mara, 2010). The magnitude and the significance of the indirect effects, that is of the standardized effects that were mediated by the intervening variables, was obtained using the MODEL CONSTRAINT command. Specifically, the indirect effects of L1-MACH1 and L2-MACH1 on L1-MACH4 were estimated by multiplying the effect of L1-MSC4 on the corresponding variable with the effect of the latter on L1-MSC4.

4.4.5 Summary of the Methodology for Study 2

For the purposes of Study 2, I used the same data sample as that in Study 1a, which referred to English primary year one and year four students. After treating missing data in the appropriate way and deriving the required variables, I proceeded with my statistical analyses. For the purposes of Study 2a, this involved the assessment of the compositional effect of school average achievement at the end of year one (L2-MACH1) on (i) students' self-concept at the end of year one (L1-MSC1) and (ii) students' self-concept at the end of year four (L1-MSC4). For both analyses, I used multilevel structural equation models from the 2x2 taxonomy. As an extension to these models and bridging Study 2a with Study 1a, in Study 2b, I tested whether mathematics self-concept at the end of year four (L1-MSC4) mediated the negative effect of school-average achievement at the end of year one (L2-MACH1) on self-concept at the end of year four (L1-MSC4).

4.5 Results for Study 2a: The BFLPE in Early Primary Years

4.5.1 Results for Research Hypothesis 2a.1: The Compositional Effect of Year One Mathematics Achievement on (i) Year One and (ii) Year Four Mathematics Self-Concept Using the Doubly Manifest Approach

In two separate analyses the magnitude and direction of the compositional effect of school average mathematics achievement at the end of year one on (i) students' self-concept in mathematics in year one and (ii) students' self-concept in year four were assessed. The approach followed was conventional multilevel modelling, namely the doubly manifest approach.

Most research on the prevalence of the BFLPE (see section 2.1.5) has been based mostly on high school students and there has been almost no research based on early primary school years. In assessing the magnitude of the big-fish-little-pond-effect for year one students (five-year-olds) and year four students (eight-year-olds to nine-year-olds) the expectation was that the BFLPE would be verified with both year groups. Note that the two distinct analyses referred to the same group of students. Indeed, the BFLPE hypothesis was supported for both year one and year four; the results are displayed in the first row of Table 4.2. A small and marginally significant negative effect of school average prior achievement was observed with year one data (β_{com} = -.032, se =.016, ES = -.03). A stronger effect was detected using year four data (β_{com} = -.190, se = .025, ES = -.244). Although these effects could only be characterized as weak based on the current thresholds in the existing literature (see Cohen, 1988; 1992); still they are indicative of the prevalence of BFLPEs with both year one and year four data. The within-group effect, that is indicative of the strength of the relationship between an individual's mathematics achievement and mathematics self-concept in years one and four, was also found to be stronger for year four students (β_{within} = .102, se = .007, ES = .241) as compared to year one students (β_{within} = .061, se .005= ES = .217). This is also in line with previous research (e.g. Guay, Marsh and Boivin, 2003).

4.5.2 Results for Research Hypothesis 2a.2-Research Hypothesis 2a.4: Using Partial and Full Correction Approaches for the Investigation of the BFLPE

This set of research hypotheses was mainly concerned with the impact of correcting for measurement and sampling error bias in the obtained BFLPE estimates and associated standard errors. To this end, I applied the set of all four partial and full correction approaches from the 2x2 taxonomy to both year one and year four data.

The results are again displayed in Table 4.2. Each row of this table corresponds to a different approach of the 2x2 taxonomy (see Table 2.1 in the literature review chapter for a reference to the four models). I guide the reader from the top row, which corresponds to the obtained effects (within-group effect and compositional effect) using the doubly manifest approach, through the last row, where the findings using the doubly latent approach are displayed. I remind the reader that when comparing the obtained effects across the four models, emphasis should be placed on the effect sizes - that are on a comparable metric - rather than on the unstandardized estimates.

4.5.2.1 The impact of adjustments for measurement and sampling error on the compositional effect of school average achievement at the end of year one on self-concept in year four

Adjustments for sampling error (manifest latent approach) led to a more negative BFLPE compared to that obtained with the doubly manifest approach (β_{com} = -.224, se = .029, ES = -.260; compare this estimate with that reported with Research Hypothesis 2a.1). Further adjustments for measurement error in the individual-level mathematics achievement and the school-level aggregate resulted in more negative BFLPEs (β_{com} = -.170, se = .022, ES = -.288). An even more negative effect was observed when the doubly latent approach was applied to the data (β_{com} = -.198, se = .025, ES = -.304). In summary my findings support, my hypothesis that the negative compositional effect of school average achievement at the end of

year one on self-concept at the end of year four became more negative after adjustments for either or both measurement and sampling error. At the same time, the compositional effect estimates obtained using the four distinct models of the 2x2 taxonomy are not substantially different from each other, this reflects the fact that the data underlying my analysis were of high reliability both in terms of measurement error and in terms of sampling error (see also section 6.2 in the Discussion chapter).

4.5.2.2 The impact of adjustments for measurement and sampling error on the compositional effect of school average achievement at the end of year one on self-concept in year one

In applying the set of the four partial and full correction approaches from the 2x2 taxonomy to year one data, the derived BFLPE estimates became more negative (see Table 4.2) – in line with my previous analyses and derived conclusions. The doubly latent approach retrieved the most negative compositional effect.

4.5.2.3 The impact of adjustments for measurement error and sampling error on within-group effects

While in the BFLPE paradigm the focus is mainly on compositional effects of school average achievement on students' self-concept, within group effects of individual achievement on self-concept can also be interesting. From a methodological point of view we can again observe that while adjustments for sampling error have generally no impact on the within-group effect estimates, adjustment for measurement error led to stronger within-group effects. This is in line with the observations that we made in Study 1a (see the comments on the impact of measurement and sampling error adjustments on within-group effects made in section 3.3.3 of Study 1a). What is interesting to note in relation to the present analysis is that the within-group estimate obtained with the latent manifest approach and the doubly latent approach are essentially the same (see Table 4.2).

This is due to the fact the two approaches differ only in the assumptions that they make on the prevalence of sampling error in the educational data, an issue irrelevant in the estimation of the within-group effects.

4.5.2.4 The impact of adjustments for measurement and sampling error on standard error estimates

Research hypothesis 2a.4 referred to the trade-offs in bias and accuracy across the four models of the 2x2 taxonomy. It is true that correcting for unreliability due to sampling error and measurement error in my analysis introduced variability in the estimation of both the within and the compositional effect (see Table 4.2) – the standard errors with the partial correction and full correction approach were somewhat larger than those with the doubly manifest approach.

4.6 Results for Study 2b: Academic Self-Concept as Mediator of the Negative Compositional Effects of School Average Achievement on Students' Subsequent Self-Concept and Students' Mathematics Achievement

4.6.1 Results for Research Hypotheses 2b.1-2b.2: The Growth of BFLPEs over the First Four Years of Primary Schooling

In Study 2a, I concluded that there was a negative and significant compositional effect of L2-ACH1 on L1-MSC4. In the present analysis I extended this model, investigating whether this negative effect of L2-MACH1 on L1-MSC4 (see results for Study 2a) was still evident after adjustments for L1-MSC1. My rationale was that, if the negative effect of L2-MACH1 on L1-MSC4 was entirely mediated by L1-MSC1, that is, if there was no direct effect, then this would mean that BFLPEs were stable over the first four primary school years: the BFLPE effect on L1-MSC4 was neither larger nor smaller than that on L1-MSC1. A negative direct effect of L2-MACH1 on L1-MSC4 would mean that BFLPEs increased over time. A positive direct effect of L2-MACH1 on L1-MSC4 would suggest that the BFLPEs decreased over time. If the BFLE

were limited to L1-MSC1, this would contradict the notion that seems to pervade existing literature (see section 2.1.5 in the literature review) namely that the negative effect of school average achievement is stable and grows more negative with time. It would also detract some of the practical importance of the BFLPE research, as it would suggest that the impact of the excellency of the schools' intake on students' self-concept`, as determined by the average achievement of the students when entering the school, only lasts temporarily without any long-term implications.

In comparing the estimates of the compositional effect of school average achievement in mathematics at the end of year one (L2-MACH1) on individual self-concept in year one (L1-MSC1) and the same effect on self-concept in year four (L1-MSC4; see Table 4.2) the latter is clearly more negative for all models of the 2x2 taxonomy. Importantly (see Figure 4.1; see also Table 4.3), while the indirect effect of L2-MACH1 on L1-MSC4 via L1-MSC1 is statistically significant (β_{com} = -.032, se =.016), the direct effect is even more negative and highly significant (β_{com} = -.307, se =.038). Therefore, it is clear that the negative effects of school average achievement on self-concept in year four persist even after adjustments for self-concept in year one: new big-fish-little-pond effects manifest during the first four years of schooling as a result of the way in which students are distributed across schools in year one.

4.6.2 Results for Research Hypothesis 2b.3: Is the Negative Compositional Effect of School Average Achievement on Subsequent Achievement (Study 1a) Mediated by the BFLPE?

Juxtaposing the negative compositional effects on self-concept detected in Study 2a with the findings of Study 1a (negative compositional effect of school-average achievement on subsequent achievement) an immediate question that arises is the extent to which the two are inter-related. In addition, considering the compelling evidence that exists in the literature suggesting that academic self-concept mediates many of the negative effects of school-average

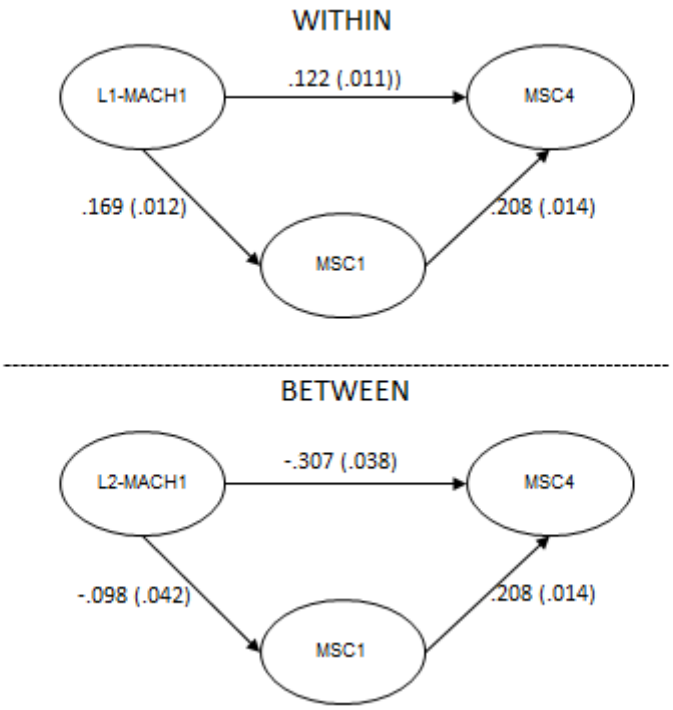
achievement on subsequent educational outcomes (see section 2.1.6 in the literature review) then research hypothesis 2b.3 (academic self-concept mediates the negative compositional effects on mathematics achievement) can justifiably be made. Indeed, my findings (see Table 4.4 and also Figure 4.2) support this hypothesis: the indirect effect of L2-MACH1 on L1-MACH4 (β_{com_ind} = -.030, *se* .004) via L1-MSC4 was found to be statistically significant. The direct effects of L2-MACH1 on L1-MACH4 were also negative and statistically significant (β_{com_dir} = -.111, *se* .042) so that the mediation analysis provided evidence for partial (rather than total; see section where I explain the difference) mediation of L1-MSC4 on the negative compositional effects of L2-MACH1 on L1-MACH4.

Table 4.2:The Big Fish Little Pond Effect for year one and year four

	Academic self-concept in year one on academic achievement in year one				Academic self-concept in year four on academic achievement in year one			
	β_{within}		β_{com}		β_{within}		β_{com}	
	Estimate ($ES_{\beta_{within}}$)	*SE*	Estimate ($ES_{\beta_{cont}}$)	*SE*	Estimate ($ES_{\beta_{within}}$)	*SE*	Estimate ($ES_{\beta_{cont}}$)	*SE*
Doubly Manifest	.061 *(.217)*	.005 *(.016)*	-.032 *(-.03)*	.016 *(.015)*	.102 *(.241)*	.007 *(.016)*	-.190 *(-.244)*	.025 *(.032)*
Manifest Latent	.061 *(.221)*	.005 *(.016)*	-.041 *(-.034)*	.019 *(.016)*	.102 *(.245)*	.007 *(.016)*	-.224 *(-.260)*	.029 *(.034)*
Latent Manifest	.057 *(.294)*	.005 *(.024)*	-.024 *(-.034)*	.013 *(.018)*	.090 *(.270)*	.006 *(.019)*	-.170 *(-.288)*	.022 *(.037)*
Doubly Latent	.057 *(.299)*	.005 *(.024)*	-.030 *(-.038)*	.014 *(.018)*	.090 *(.274)*	.006 *(.019)*	-.198 *(-.304)*	.025 *(.039)*

Note. The parameter β_{within} denotes the within group effect of the outcome variable (mathematics self-concept in Year four) on the individual-level predictor (mathematics achievement in Year one) while β_{cont} denotes the effect of school average achievement in Year one on mathematics self-concept in Year four. SE denotes the standard error of the parameter estimate and of the corresponding effect size; ES denotes the Effect Size estimate for the corresponding effect.

Figure 4.1: Testing the stability of the BFLPE during the first four years of primary school



Note. In the above path diagram L1-MACH1 is individual mathematics achievement in year one, L2-MACH1 is school average achievement in year one, MSC1 is individual self-concept with respects to mathematics in year one and MSC4 is individual self-concept with respect to mathematics in year four.

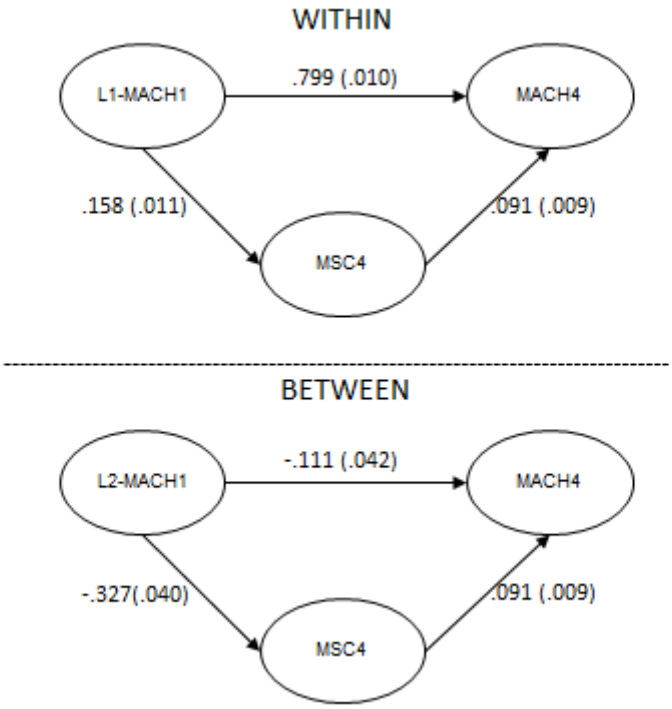
Table 4.3:The negative compositional effect of year one school average achievement on year four self-concept, after adjustments for year one self-concept

Covariate considered	**Direct effects**	**Total Indirect effects**	**Total effects**
L1-MACH1	.122***	.035*	.157***
L2-MACH1	-.307***	-.021*	-.329***

Note. L1-MACH1 in individual-level mathematics achievement at the end of year one, L1-MSC1 is individual-level mathematics self-concept in year one, L2-MAC1 is school-level mathematics achievement at the end of year four. The direct, total indirect and total effects refer to the effects of the corresponding variable on mathematics self-concept at the end of year four (L1-MSC4). The mediator variable considered is academic self-concept in relation to mathematics at the end of year one.

*p<.05, **p<.01, ***p<.001

Figure 4.2: Testing whether negative compositional effects of school average mathematics achievement on subsequent mathematics achievement are mediated by mathematics self-concept



Note. In the above path diagram L1-MACH1 is individual mathematics achievement in year one, L2-MACH1 is school average achievement in year one, MSC4 is individual self-concept with respects to mathematics in year four and MACH4 is individual mathematics achievement in year four.

Table 4.4:Total, direct and indirect effects of year one individual achievement and school average achievement on students' year four mathematics achievement

Covariate considered	Direct effects	Total Indirect effects	Total effects
L1-MACH1	.8***	.014***	.814***
L2-MACH1	-.111***	-.030***	-.141***

Note. L1-MACH1 in individual-level mathematics achievement at the end of year one, L2-MAC1 is school-level mathematics achievement at the end of year four. The direct, total indirect and total effects refer to the effects of the corresponding variable on mathematics achievement at the end of year four. The mediator variable considered is academic self-concept in relation to mathematics at the end of year four. *p<.05, **p<.01, ***p<.001

4.7 Study 2: An Overview of the Findings

In general, my research hypotheses for both components (Study 2a and Study 2b) of Study 2 were verified. The compositional effect of school average achievement at the end of year one on students' self-concept at the end of year four was weak negative and significant, establishing the presence of big-fish-little-pond-effects with the data used in Study 1a. Big-fish-little-pond-effects were also verified with year one data, albeit weak. Adjustments for measurement error with or without further adjustments for sampling error through the models of the 2x2 taxonomy led to more negative big-fish-little-pond-effects, also consistent with my hypotheses. Then, in testing my first mediation hypothesis I found that the negative compositional effect of school average achievement at the end of year one on students' self-concept at the end of year four persisted even after adjustments for year one self-concept in the BFLPE model. Thus, there were additional big-fish-little-pond-effects through the first four years of primary schooling. Lastly, in testing my second mediation hypothesis, namely whether self-concept mediated the negative compositional effect of school average achievement in year one on students' self-concept in year four, I demonstrated partial mediation in the relevant model.

Chapter 5: An Application of the Regression Discontinuity Approach to English TIMSS 95 data: Addressing the Issue of Measurement Error Bias and Exploring the Possibility to Investigate Potential Effects of School Composition (Study 3)

5.1 Introduction

In the first study of my thesis, Study 1, the main focus was on school compositional effects and on how relevant estimates alter when measurement error and sampling error bias is corrected for. Multilevel compositional analyses models from the Marsh, Lüdtke et al. (2009) 2x2 taxonomy were used throughout the analysis. In this study, Study 3, I look at the same substantive issue, namely the potential effects of the school composition, using an entirely different methodological approach, in an extension of the classic Regression Discontinuity (RD) approach. I seek to investigate the extent to which school composition, as quantified by aggregates of student achievement, is correlated with added-year effects, rather than with schools' value added scores (Study 1).

The RD approach, applied in any relevant research setting, allows the estimation of the impact of a treatment accurately, provided the precise criterion of assignment to treatment and control group is known (Shadish, Cook and Cambell, 2002; Rossi et al., 2004). In educational research, the RD approach (see section 2.4.2 in the literature review) can be used as an alternative approach for studying school effectiveness (Cook, 2008; Kyriakides and Luyten, 2009; Luyten, 2006; Luyten, Peschar and Coe, 2008, Luyten, Tymms and Jones, 2009) once the main assumptions underpinning the approach are fulfilled (see section 2.4.5 for the assumptions underpinning the RD approach). Specifically, for educational systems in which students begin formal education strictly based on an age cut-off, the RD approach allows the estimation of the

absolute effect of schooling, that is, the effect of an extra year of schooling, and the effect of chronological age on achievement simultaneously.

In Study 3, I begin by applying the Regression Discontinuity (RD) approach to investigate the absolute effect of schooling in England (see also Cliffordson, 2010; Luyten, 2006), using primary school mathematics achievement data (years four and five) from the Third International Mathematics and Science Study 95 (Study 3a). I also investigate the absolute effect of schooling with secondary school mathematics achievement data (years eight and nine) from the same database (Study 3b). In both studies, I employ multilevel models; with the use of multilevel modelling it is possible to investigate differences across schools in their estimated effects and thus obtain relative school effects as well as absolute schooling effects.

The fact that the RD approach does not require adjustments for prior achievement or any other background characteristics is of considerable practical value (see section 2.5.6 in the literature review on the advantages of the RD approach): added-year effects have been claimed to reduce bias in estimating school outcomes due to differences in student background (see for example Heck and Moriyama, 2010). I explore this possibility in my analysis controlling for the main effects of a range of student-level variables in my models and observing the impact on the estimates of the absolute effect of schooling. These are chosen strategically to resemble the set of variables that have been used until recently in value added models, especially in the E-CVA models that were until recently being used in England (see Ray, 2006; also see "Value Added Models of educational effectiveness and relative school effects" in section 2.3.5 of the literature review).

As a complementary analysis, I investigate the extent to which each of these variables moderates the absolute effect of schooling; a significant interaction would suggest differential school effectiveness for different groups of students. Differential school effectiveness is a term that is used in educational effectiveness literature (see, for instance, Dearden, Mickelwright and Vignoles, 2011) to denote the possibility that some schools may be differentially effective, for

example by being more effective for high performing pupils and less effective for low performing pupils: “Do some schools do better in assisting particular types of pupils (the above average, for example), to obtain examination success than they do with others, and vice versa?” (Gray, Jesson and Sime, 1990).

Moreover, using the multilevel RD model, I consider the extent to which school-level variables can be used to explain the variability across schools in their absolute effects. To be precise, I explore the potential of using the RD approach as a tool to assess the effect of variables related to the schools’ composition (and, specifically, school average achievement) on students’ outcomes. A relevant research question in the RD framework would be: “To what extent can the variability across schools in their absolute effects be explained by variables measuring school composition (e.g. school average achievement?)”. In this respect, one of the main purposes of Study 3 is to demonstrate the strengths of the RD approach over conventional compositional models in assessing the impact of school composition on students’ outcomes.

In addition, in Study 3, I also consider the impact of adjustments of measurement error in students’ outcome scores on the RD estimates, both when variables related to the school composition are used in the model and when they are not. When school-level aggregates are used in the analysis, the impact of simultaneous adjustments for both measurement error and sampling error on estimation is also investigated. In pursuit of this aim, as well as using the conventional multilevel modelling approach, I fit RD models using multiple indicators for mathematics achievement – these are defined by the different content areas as specified in TIMSS-95 (see following section 5.3.2 and Table 5.3 where these content areas are given). Multilevel structural equation models, integrated within the RD framework, are used throughout my analysis and sampling error is corrected for assuming latent aggregation (see Lüdtke et al., 2008).

I begin this chapter with the main research questions and research hypotheses related to primary school TIMSS 95 (Study 3a). The same set of research questions and research

hypothesis are addressed in relation to secondary school data (Study 3b). The methodology followed was common for both studies and hence it is presented in one single section. The findings are presented in two separate sections, one for primary school data and the other for secondary school data.

5.2 Research Questions and Research Hypotheses

In this section I give the main research questions and research hypotheses underlying Study 3. I begin with research hypotheses underlying the application of the RD model to primary school data to assess the absolute effect of schooling as well as differences across schools in their absolute effects ("*The absolute effect of schooling and the effect of chronological age on students' mathematics achievement"*). These are based on empirical findings from previous studies that also used a similar approach to investigate absolute schooling effects (see Luyten, 2006). Then in the sections that follow, I present research questions which are addressed simultaneously for primary (Study 3a) as well as for secondary (Study 3b) school data. They are organised into three sets of analyses: The first is concerned with the impact of adjustments for the main effects of variables related to the students' background on regression discontinuity estimates of the absolute effect of schooling and relative differences across schools in these effects. ("*The effect of adjustments for student background variables on added-year effects/ Investigating differential school effectiveness"*). The second aims to explain between school differences in their absolute effects using school average achievement as a level 2 covariate in the slope of the absolute effect of schooling ("*Examining relationships among schools' composition and added-year effects"*). Lastly, the third investigates effects of school composition (average achievement and socio-economic status) on added-year effects employing multilevel structural equation models, rather than multilevel models. In this way, adjustments can be made for (i) measurement error in students' achievement, the dependent variables in RD models and for (ii) level-2 measurement error in school-level variables as well as for (iii) sampling error in aggregating individual-level variables to form the school-level measures. The

same set of research questions and research hypotheses are addressed with secondary school data ("*Applying the RD approach to secondary school TIMSS-95 data").*

5.2.1 The Absolute Effect of Schooling and the Effect of Chronological Age on Students' Mathematics Achievement

In the first part of Study 2 I replicate previous studies (see Luyten, 2006) concerned with the application of the regression discontinuity approach to investigate absolute schooling effects with primary TIMSS 95 mathematics achievement data. Based on empirical results, I form research hypotheses as follows:

Research Hypothesis 3a.1: The effect of chronological age on achievement is expected to be positive. The estimated effect of one year of schooling on students' mathematics achievement is also predicted to be positive; albeit smaller than the effect of age.

Older students in a year group tend to attain higher scores: certainly the older students in the class are more mature and, particularly in the initial stages of schooling, they perform generally better than the younger students (Bedard and Dhuey, 2006). A wide range of studies suggest that one year of schooling also exerts positive effects on achievement. This is sometimes reported to be about twice as strong as the effect of one chronological year (Cahan and Cohen, 1989; Cahan and Davis, 1987). Nevertheless, especially with the English primary TIMSS-95 data, the effect of schooling was found to be smaller than the effect of chronological age (Luyten, 2006).

Moreover, Cahan and Cohen (1989) claim that the within grade increase in performance can be approximated satisfactorily by short linear segments of two years. With TIMSS-95 data quadratic and cubic relationships were tested; these were found non-significant (Luyten, 2006).

Research Hypothesis 3a.2: The linear model that describes the relationship between age and achievement has the same slope both in the lower (year four) and in the higher (year five) grade.

The regression discontinuity approach in my study is applied in a multilevel linear regression modelling framework. I investigate the variability between schools in their absolute effects; thus combining important aspects of the RD (with a focus on absolute effects of schooling) and compositional models like those used in Study 1 (with a focus on school-to-school variation; the relative effects of individual schools and their relation to school composition variables).

Research Hypothesis 3a.3: There exists significant school-to-school variation in the absolute effect of schooling across schools.

5.2.2 The Effect of Adjustments for Student Background Variables on Added-Year Effects/ Investigating Differential School Effectiveness

As a second step in my analysis, I control for a range of variables related to the students' background. In fact, the RD approach does not require adjustments for background variables once the main assumptions of the approach are fulfilled (see also section 6.6.3 in the discussion chapter): omitted variable bias is not a problem with regression discontinuity estimates of absolute schooling effects. Including additional variables in the basic models (see section 2.5.3) is expected to affect the estimates of the absolute schooling effects minimally because students from the distinct grades should be homogeneous in relation to background characteristics at the cut off– a critical assumption of the RD approach.

Research Hypothesis 3a.4: Adjustments for the main effects of student background variables in the regression discontinuity models are expected to leave unaltered the RD estimates of the absolute effect of schooling.

Although the main effects of additional independent variables are not the main focus of my analysis, it is still relevant to examine the significance of their interactions with the grade level effect. The following research question derives from the differential effectiveness literature:

Research Question 3a.5: Is the effect of absolute effect of schooling differentially effective for students that come from different backgrounds (e.g. ethnicity, socio-economic status, gender)?

Importantly, although of interest in its own right, such differences have no effect on the appropriateness of the RD approach so long as these individual student background characteristics are not systematically different for students who are at the cut-point (the oldest students in year four and the younger students in year five). This assumption is likely to be valid as long as age is the sole determinant when students start school.

5.2.3 Examining Relationships among Schools' Composition and Added-Year Effects

Evidence for between school variance in the absolute effect of schooling suggests that not all schools influence their students' attainment through one year of schooling in the same way. This residual variance in added-year effects is the measure of school effectiveness in regression discontinuity designs. Here, I use the RD approach to address debates on the relationship between schools' composition and school effectiveness.

In the analysis I present here I use the average achievement across all students in the school to explain between school differences in their absolute effects; this is estimated by aggregating achievement across students from both the lower and the higher grade. This school-level aggregate has repeatedly been used in the literature to capture effects of the school's composition (Willms, 1985b; Televantou, Marsh, Kyriakides, Nagengast and Fletcher, in press; Marsh, Nagengast, Fletcher and Televantou, 2011). Nevertheless their effects on students' outcomes and their relationship with educational effectiveness measures have only been investigated in the value added modelling framework which uses compositional analysis. One of the main limitations underlying this approach, and for which it has often been criticized is that, in order to be able to capture unbiased estimates, all potential student background variables are required (see section 2.5.5 in the literature review) and it is never possible to argue that all the

important background variables have been included. For regression discontinuity designs this critical limitation is no longer a requirement (see Research Hypothesis 2a.5).

Research Question 3a.6: Do school-level variables relevant to school's composition (school's socioeconomic status, average attainment of students) have a significant effect on the size of the gap for schools?

5.2.4 Integrating Multilevel Structural Equation Models with Regression Discontinuity Designs

The originality of Study 3 lies in the fact that not only does it employ the multilevel modelling framework and consider the impact of school composition variables in explaining between school differences in their absolute effects, but also in the fact that it is concerned with the impact of measurement error on RD estimates.

The basic RD model involves regressing students' mathematics achievement on age and year of schooling. Three variables are basically involved in RD models (see section 2.4.3 where I give the equation for the RD model): students' achievement, age and a dummy variable denoting students' year group: lower or higher. Students' age does not involve substantial amounts of measurement error and is assumed to be measured without error. The classification of students in lower and higher grade might involve misclassification error, but these are likely to be minimal as well (but also see Goldstein, Kounali and Robinson, 2008 for a different approach when misclassification might be substantial). In my analysis I was concerned only with the prevalence of measurement error in students' measured achievement - the criterion in RD models. Statistical theory suggests that adjusting for random error in the response variable of regression models should not alter the expected value of the estimates (see also Ferrão and Goldstein, 2008; see Appendix B, section B.3 for the relevant mathematical derivations). Nevertheless it can result in larger standard errors, a consequence of the unreliable measurement of the criterion. Crucially, in relation to our data, which were of high reliability coming from the

TIMSS 95 database, it was expected that controlling for measurement error in students' mathematics achievement should not have substantial impact on both RD estimates and associated standard errors.

Research Hypothesis 3a.7: Adjustments for multilevel measurement error in the mathematics scores of the students, the criterion in RD models, should have little or no impact on the unstandardized regression discontinuity estimates and associated standard errors

In a second set of analyses, I used multilevel structural equation models to investigate effects of school composition (school average achievement) on the residual variance of the effect of one extra year of schooling (see section *5.4.3*).

Research Question 3a.8: What is the magnitude of the effect of school average achievement, as used to explain variability across schools in their absolute effects, when adjustments are made in the RD model for measurement error at level-1 and level-2 and sampling error in the school-level aggregate?

5.2.5 Applying the RD Approach to Secondary School TIMSS-95 data

This is the first study to use the RD approach and the multilevel modelling framework with secondary school TIMSS-95 data: By addressing the same set of research hypotheses for students in higher school years, descriptive comparisons can be made in relation to the effects of schooling and age on students' cognitive development across different phases of schooling (primary vs secondary school). The evidence from previous literature is that both the effect of schooling and the effect of age will be smaller as students grow older (Bedard and Dhuey, 2006; Crone and Whitehurst, 1999). For instance, in relation to the absolute effect of schooling, a year of schooling for students in the fourth or fifth year is almost a quarter of the total education received. For an older student, a year of schooling is not such a substantial part of his total time in school.

In addition to the set of research questions investigated for Study3a, and which are identical to Study 3b, the following research hypothesis was therefore addressed:

Research Hypothesis 3b.9: The estimated effects of schooling and age on achievement are larger for primary school data as compared to secondary school data.

The Research Questions and Research Hypothesis chapter is the basis for the Results chapter where the main results of the present study are presented. Throughout the Results chapter, the notation "3b.x" is used to refer to the research question or the research hypothesis addressed (as opposed to the notation "3a.x" used for primary school data).

5.3 Methodology

The methodological approach was common for Study 3a and Study 3b. A detailed description of the methods and procedures followed is given in the next three sections. First the data samples involved are described. Subsequently, I elaborate on the set of variables used throughout modelling. Lastly, the distinct steps in the analyses are explained.

5.3.1 Data Samples

5.3.1.1 Use of both primary and secondary school data

In Study 3, I used English primary and secondary school mathematics achievement data from the "Third International Mathematics and Science Study" conducted in 1995 (TIMSS 95). In TIMSS 95 framework, the population of primary school students is defined for each country as the population of students enrolled in the two adjacent grades that contain the largest proportion of nine-year-old students at the time of testing. Thus, the English primary school data used for this study involve years four and five. In the same way, the population of secondary school students consists of students enrolled in the two adjacent grades containing the largest

proportion of thirteen-year-old students at the time of testing and. For England these involve years eight and nine.

5.3.1.2 The use of the Third International Mathematics and Science Study (TIMSS 1995)

TIMSS 95 was conducted across more than forty countries and tested more than half a million students in mathematics and science. Several subsequent studies followed up – these took place in 1999, 2003, 2007 and 2011. The objective of these surveys is to provide policy makers, educators, researchers and practitioners with internationally comparable information on the measured achievement and learning context of the students in the two subjects of science and mathematics (Luyten and Veldkamp, 2011).

Although more recent TIMSS studies have been conducted, TIMSS 95 is the only test administered involving two adjacent grades at the primary and the secondary level. The remaining TIMSS studies assessed only one primary and one secondary school grade. Therefore, more recent TIMSS data cannot be used to apply the RD approach.

5.3.1.3 Misclassification: Percentage of delayed/accelerated cases in the samples

The RD design is based on strong assumptions (see also section 2.5.5 in the literature review): It is assumed that school entry is based on age only, but in reality some individuals may start school later or earlier depending on their age and intellectual development. When the percentage of non-normal aged children does not exceed 5%, students with no normal age can be safely excluded in a list-wise manner from the data (Cliffordson and Gustaffson, 2011; Luyten, 2006; Luyten et al., 2009; Judd and Kennedy, 1981; Trochim, 1984; Shadish et al., 2002).

The percentages of students with accelerated/ normal/ delayed school careers involved in the TIMSS 95 primary and secondary school samples are displayed in Table 5.1 and Table 5.2. The mean mathematics achievement (based on the overall scale scores used for the purpose of the present study; see section 5.3.2) and standard deviation for each one of these groups of students is also given. In the lower grades the mean achievement of accelerated students was on average higher than the mean achievement of students who were delayed or those who followed a normal schooling career, being classified in lower and higher grades according to their age. This phenomenon could be observed both with primary and secondary school data. What is also interesting is that for primary school, the delayed students in the lower grade (year four) actually did better than those who followed a normal school career. Moreover, the accelerated students in the higher grade (year five) did better than those who were classified as normal in the time they started school.

Only students with normal school careers were considered in the analysis: The cases with accelerated or delayed school careers in the data were a negligible number – of much smaller proportion than the nominal five per cent cut off – and could thus be safely deleted from the analysis.

Table 5.1: Number of cases, mean and standard deviation for mathematics achievement scores for accelerated, normal-aged and delayed students for years four and five[1].

Year Group	Number of students	Number of schools	Number of accelerated-normal aged-delayed	Missing cases	Accelerated		Normal		Delayed		Missing	
					Mean	Standard Deviation	Mean	Standard Deviation[2]	Mean	Standard deviation	Mean	Standard deviation
Four	3056	126	22(.007%) - 2943(96.3%) - 55(.018%)	36	448.86	107.85	421.27	98.17	423.24	95.52	454.03	85.18
Five	3126	127	25(.008%) - 3017(96.5%) - 4(.001%)	80	504.91	115.09	481.91	95.78	552.17	95.19	478.87	81.24

[1,2] Note. [1] In the data there exist 6182 students nested in 134 schools. The average number of students within each school is equal to 46, ranging from 18 to 68.[2] The pooled standard deviation across year four and year five for the normal aged students is 96.86.

Table 5.2: Number of cases, mean and standard deviation for mathematics achievement scores for accelerated, normal-aged and delayed students for years eight and nine[1].

Year group	Number of students	Number of schools	Number of accelerated-normal aged-delayed	Missing Cases	Accelerated		Normal		Delayed	
					Mean	Standard Deviation	Mean	Standard Deviation[2]	Mean	Standard deviation
Eight	1803	123	16 (.89%) – 1762(97.73%)– 25 (1.39)	-	528.2481	80.54724	496.6896	90.15203	448.7912	95.51836
Nine	1776	122	23 (1.3%) – 1735 (97.7%)– 18 (1.0135)	-	476.4213	91.29237	499.4100	85.80833	466.6444	97.46344

[1,2]*Note.* [1]In the dataset there exist in total 3059 students and the total number of schools is 128. Each school has, on average 28 students, with numbers ranging from 21 to 32. [2] The pooled standard deviation across year eight and year nine for the normal aged students is 88.

5.3.2 Variables

In this section I provide a description of the variables that were used throughout the analyses: the students' achievement measures in mathematics used with multilevel manifest models and multilevel structural equation models, the student background variables incorporated to test the impact of adjustments of level 1 covariates on RD estimates and the school-level covariates quantifying the effects of the schools' composition on added-year effects.

5.3.2.1 Achievement measures

For the present analysis two sets of achievement measures were required: Scale scores for the students' overall mathematics achievement – these were incorporated through the different analyses for (research hypothesis 3x.1- research question 3x.7).Whenever achievement was conceptualized as a latent variable (research hypothesis 3x.8- research question 3x.9), multiple indicators were used to control for measurement error.

The TIMSS international database contains several student-level achievement scores that were computed at different stages of the study to serve specific purposes. For the purposes of this study the new scaled scores that were released in 1999 were used. These scores were computed using a different psychometric model (three parameter model) from the one originally used (one parameter model) for the Rasch scores released in 1995 and are set on the scale that will be used to measure trends in mathematics and science in future TIMSS assessments (http://timss.bc.edu/timss1995i/Database.html#DBnewScaleScores). Most importantly, in the most recently released database, scales can be found not only for the total mathematics achievement score of the students but also for the different content areas that were assessed. Indeed, one of the main advantages of the TIMSS database, which has not yet been fully exploited, is that not only measures of the mathematics achievement of the students, but also measures of the attainment of the students in different sub-areas of learning are provided.

For primary school data (years four and five), four different content areas exist: Whole numbers, fractions and proportionality, geometry and measurement and data replication, analysis and probability. For secondary school data (years eight and nine), five different areas are assessed: Fractions and number sense, geometry, algebra, data replication, analysis and probability and measurement. The scales on the different content areas are used as multiple indicators to control for measurement error in my models.

In Table 5.3 the mean and standard deviation for each content area is displayed for primary and secondary school data, respectively.

Table 5.3: Mean and Standard Deviation for the different content areas in TIMSS 1995 primary and secondary school data

Content area	Lower Grade		Upper Grade	
	Mean	Standard deviation	Mean	Standard deviation
Primary school data				
Whole numbers	399.9	98.7	460.3	92.6
Fractions and Proportionality	429.1	91.9	485.7	92.1
Geometry and Measurement	454.9	96.9	509.9	94.9
Data Rep., Analysis and Probability	464.4	102.4	518.0	97.3
Secondary school data				
Fractions and Number sense	474.0	90.1	497.0	83.9
Geometry	460.2	83.8	486.5	82.4
Algebra	464.9	88.2	495.9	82.8
Data Rep., Analysis and probability	492.2	95.1	512.2	92.8
Measurement	477.5	90.1	505.4	82.9

5.3.2.2 Student background variables

In my models I adjust for a number of student-level variables in addition to age and grade - the basic covariates included in regression discontinuity models. These were chosen based on previous research (see for example Luyten, 2006) but also in an attempt to mimic covariates used in the Contextual Value Added model (CVA) model (Ray, 2006) which has been used in the UK for the assessment of the effectiveness of primary and secondary schools. In Table 5.4 , I give a list of these variables together with the percentage of missing data both for primary and secondary school data.

Gender: The first variable investigated was the students' gender. It was recoded so that girls would get a value of zero and boys a number of 1. The number of boys and girls was approximately the same in both samples - 3019 (50.65%) girls and 2941 (49.35%) boys in with primary school data and 1638 (46.84%) girls and 1859 (53.16) boys with secondary school data

Total number of people in the students' home: The total number of people in the student's home was a continuous variable ranging from 2 to 19 for primary school data (mean=4.6045; sd=1.3860) and from 1 to 15 for secondary school data (mean=4.4983, sd=1.3255).

Total number of books at the students' house: The number of books at the student's house denoted the approximate number of books in the student's house (0-10, 11-25, 26-100, 101-200, more than 200). The number of cases in each category for the two datasamples is given in the Appendix.

Ethnicity: Three binary variables were included in my models that provide some information on the student's ethnicity (Were you born in the UK? Was your mother born in the UK? Was your father born in the UK?). The student's ethnicity is one of the covariates used in the CVA model – 18 dummy variables related to the student's ethnicity and one for students with unspecified ethnicity are included in the model.

First Language: The frequency with which English – the language of the test - is spoken at home (always/almost always, sometimes, never) should give some indication whether or not English is the student's first language– a variable that is directly included in the CVA model (Is English not the student's first language?).

Home possessions: Then four dummy variables that denote home possession in relation to calculator/ computer/ study desk for own use/ dictionary are included in the model, these could provide some indication of the socioeconomic status of the students. In the CVA model two indicators aim to measure the students' status: (a) An indicator denoting whether the student is eligible for Free School Meal and (b) the Income Deprivation Affecting Children Index (IDACI).

Table 5.4: Proportion of missing data in background variables for primary and secondary school data

Variable description	Number (percentage[1]) of missing cases	
	Primary School Data	Secondary School Data
Student-level variables[2]		
Gender	0	23 (.658%)
Number of people in the student's house	354 (.059%)	45 (1.2%)
Number of books in the student's house	385 (.065%)	47 (1.2%)
Were you born in the UK?	221 (3.7%)	30 (.9%)
How often do you speak English at home?	569 (9.5%)	188 (5.4%)
Was your mother born in the UK?	851 (14.3%)	148 (4.2%)
Was your father born in the UK?	1057 (17.7%)	122 (3.5%)
Do you have a calculator at home?	266 (4.5%)	28 (.8%)
Do you have a computer at home?	266 (4.5%)	28 (.8%)
Do you have a study desk for your own use?	266 (4.5%)	28 (.8%)
Do you have a dictionary at home?	266 (4.5%)	28 (.8%)

[1,2]*Note.*[1] The percentages are given relative to the total number of cases after excluding the students with not normal age. [2] The cases with missing student-level covariates were spread around schools without any systematic pattern. Nevertheless, for primary school student-level, variables appeared to be missing for all the students in certain schools. Analysis was run both excluding and including those schools in the analysis. No systematic differences were found in the two analyses.

5.3.2.3 School composition

The primary focus in using variables related to the schools' composition lay in explaining differences between schools in added-year effects. Specifically, the average mathematics achievement of students in the school was considered in the analysis:

Primary school data: The value of the school average mathematics achievement ranged from 361.98 to 640.05 for primary school data, with a mean of 453.73 and standard deviation of 44.36.

Secondary school data: School average mathematics achievement ranged from 397.60 to 630.24 with a mean of 484.44 and a standard deviation of 49.29.

5.3.2.4 Linkage of different datasets

In order to simultaneously consider the achievement measures with the student background variables and the school composition variables, three different data files were merged: The "Student Background" file, the "School Background" file and the data file containing the new scale scores released in April, 1999. The "Student background" file contains a series of identification variables, link variables, sampling variables, achievement variables and a set of variables derived specifically for creation of international reports. The "School background" file also contains a series of school-level variables. They were completed by the principals or the administrators of the schools and provide information on the content and practices of the school. When the analysis is performed across countries, it is necessary to use both the country and the school id to merge the two files together. Nevertheless, since my analysis was limited to only the English sample, the merging was based only on the school id.

5.3.2.5 Use of plausible values in TIMSS 95

Each student participating in TIMSS 95 mathematics assessment was administered only a fraction of items within each of the content areas because of time constraints. In order to compensate for that and achieve reliable results of the students' scores, multiple imputation was used assigning five different "plausible values" to each student. A plausible value is an estimate of how the individual student would have performed on a test that included all the items in the assessment. It is based on the responses of those items that were actually answered by the student and the performance of the students with similar characteristics. I follow the same strategy as the one followed in the TIMSS international reports in achievement, using the first plausible value for the overall achievement test as well as each of the content areas: the reliability of the achievement measures, as indicated by the inter correlations between the five plausible values is sufficiently high so that the imputation error can be ignored.

5.3.2.6 Use of weights

TIMSS used a two-stage stratified cluster sample design. In the first stage a sample of schools within each country was selected and, in the second stage, a sample of classes in the sampled schools was selected. In order to make correct inferences when estimating the population parameters, the sampling design has to be taken into account: The use of appropriate weights for each respondent is required, ensuring that the distinct subgroups in the sample are properly and proportionally represented in the computation of population estimates. Following guidelines from previous research (Luyten, 2006; Luyten and Veldkamp, 2011) as well as the recommendations in the user guide for the TIMSS 95 database appropriate weights (labelled HOUWGT in the TIMSS database) are selected that sum to the sample size within each country. The sampling weights are used in order to be able to use the actual sample size that will be used in performing significance tests.

5.4 Statistical Analyses

Four different phases can be distinguished in the implementation of the statistical analyses, in accordance with the way in which the research questions research hypotheses were grouped together (see section 5.2). The statistical analysis described here was implemented first with primary school data and then with secondary school data.

5.4.1 The Absolute Effect of Schooling and the Effect of Chronological Age on Students' Mathematics Achievement

The first model was the "empty" random intercept model (see Snijders and Bosker, 2004) that only gave an indication of the amount of variance in the achievement scores of the students that could be attributable to schools and also served as a basis for calculating the effect size for the random effects. Then, in a series of nested models, adjustments were made (i) the effect of grade only (Model 1), (ii) for the simultaneous effects of grade and age (Model 2), (iii) for the simultaneous effects of grade and age allowing for the effect of grade to vary across schools (Model 3) and (iv) for the effect of age and grade, allowing the age effect to vary within each year group (Model 4). In this way, Model 3 was a random slopes model (Bryk and Raudebush, 2002) with a random effect for the grade level and Model 4 additionally included an interaction between grade level and age. The statistical equations that describe the models considered are given in the supplementary materials, in section B.1.1 of Appendix B.

5.4.2 The Effect of Adjustments for Student Background Variables on Added-Year Effects: Investigating Differential School Effectiveness

In a second set of analyses, each of the student-level variables considered was included in the basic regression discontinuity model one at a time. Interactions with the grade level were also controlled for in the models to test whether the schooling effect varied for students coming from different backgrounds. All significant main effects and their interactions, as detected in this

initial, exploratory analysis were subsequently included in the same model. Of special interest was the way in which the magnitude and significance of the absolute effect of schooling changed after such adjustments. The expectation was that, if the assumptions on which the RD approach was based were fulfilled, no difference should be observed in the estimated effects. In this respect, these analyses provide a test of assumptions underlying the RD approach. The statsistical equations that describe these analyses are given in section B.1.2 of Appendix B.

5.4.3 Examining the Relationship of Schools' Composition and Added-Year Effects

A significant variation in the absolute effects across schools would suggest that schools are differentially effective in their absolute effects. The extent to which these between school differences could be explained by differences in the characteristics of the body of the school was investigated controlling for school-level variables in the slope for the absolute effect of schooling, and specifically for school average achievement in mathematics. No adjustments were made for student background in the models (see Research Question 3x.7).

5.4.4 Integrating Multilevel Structural Equation Models with Regression Discontinuity Designs

Multiple indicators were incorporated in the last set of analysis to control for measurement error in students' achievement. The scores of the students in the distinct content areas (see Table 5.3) were treated as multiple indicators and students' mathematics achievement – the criterion in RD models was conceptualized as a latent variable, possibly measured with some error.

When testing the extent to which of school-level average achievement could explain variance in added-year effects, I controlled for measurement error using multiple indicators, and sampling error, assuming latent aggregation. To explain the between school differences in the effect of schooling, the significance of the interaction between the dummy variable denoting the year (lower or higher) and average achievement was tested (see section D.2 in Appendix D on

the Mplus code for the application of the RD approach in a multilevel structural equation modelling framework to investigate effects of school composition). Multiple indicators were used for the criterion (latent achievement) and latent aggregation was used to form the school-level indicators for average achievement. All statistical analysis was implemented in Mplus 7.

5.4.5 Effect Size Metric for the Effects of Age, Absolute Effects of Schooling and Relative Differences across Schools in their Absolute Effects

Effect size estimates for the absolute effect of schooling was computed to assist in comparisons of the importance of relevant variables across distinct analyses. Based on previous research for similar interventions and target populations (Tymms, Merrell and Henderson, 1997; Marsh, Lüdtke et al., 2009) the effect size metric used was described by the following formulae:

$$ES = \beta / SD_{pooled} \quad (5.1)$$

In equation (5.1) the value of the unstandardized regression coefficient was divided by the pooled standard deviation of the mathematics achievement scores of the students in the higher and in the lower year group. That is, SD_{pooled} was given as:

$$SD_{pooled} = \sqrt{\frac{SD_{lower}^2}{2} + \frac{SD_{higher}^2}{2}} \quad (5.2)$$

In relationship (5.2), SD_{lower} is the standard deviation of the students in the lower grade while SD_{higher} is the standard deviation of the students in the higher grade.

The same equation, multiplied by two, was used to calculate effect size measures for the effect of age (this is a continuous variable; see Tymms, Merrell and Henderson, 1997, p.112).

Similarly, denoting the school-level residual variance by σ_U , then the formula that describes the effect size for the school-level residual variance is given by the following relationship:

$ES = 2 * \sigma_U . / SD_{pooled}$ (5.3)

5.4.6 Summarizing the Methods and Procedures

Two parallel analyses are conducted as part of Study 3: One with primary school data (Study 3a) and one with secondary school data (Study 3b). The methodology followed is common for the two databases: At first, the percentage of misclassified cases in the data is identified and, given that it is sufficiently low; these cases are deleted from the analyses. Then the data files that are required to address the research questions and research hypotheses addressed (see section 5.2) are merged. The basic RD model is initially fitted to the data, allowing for random slopes of the dummy variable indicating the year of school, to investigate between-school differences in their absolute effects. Subsequently, the main effects of student-level variables are incorporated, demonstrating that no difference is observed in the absolute schooling effects when these are considered in the modelling, consistent with the assumptions underlying the RD approach. Interactions between the student background variables and the absolute effect of schooling are considered in order to investigate differential school effectiveness**.** The RD estimates of school effectiveness are robust in adjustments for student background characteristics; therefore the approach can be used as an alternative to compositional analysis to assess the impact of school composition on added-year effects for schooling. Both multilevel models and multilevel structural equation models are used to investigate such effects in my analyses. In order for the estimates to be comparable, effect size measures are also reported in addition to the unstandardized effects.

.

5.5 Results for Study 3a: Primary School Data (Years Four and Five)

5.5.1 Results for Research Hypothesis 3a.1 – Research Hypothesis 3a.3: The Absolute Effect of Schooling and the Effect of Chronological Age on Students' Mathematics Achievement

I remind the reader that the focus of research questions 3a.1-3a.4 was on estimating the simultaneous effects of one extra year of schooling and age on students' mathematics achievement. The extent to which the effects of age on achievement could be described by the same linear relationship for students in the lower grade as for students in the higher grade was investigated. Lastly, the variability of the effect of one extra year of schooling across schools was studied. The same analysis was conducted with both primary and secondary school data and the estimated effects of schooling and age for the two populations were compared.

The results of the series of nested models for primary school data are displayed in Table 5.5. This set of analysis revealed a positive and significant effect of one extra year of schooling on students' achievement (γ_{20} = .142, *se* =.047, ES = .147). This effect varied significantly across schools ($Var(U_{1j})$ = .04, *se* =.015, ES = .082) but the size of this school-to-school variation is very small. Thus, for primary school data there was evidence of absolute schooling effects as well as of relative effects across schools in their absolute effects – albeit small. The effect of age was also positive (γ_{10} = .484, *se* =.036, ES =.998).The interaction between age and grade was also found non- significant, suggesting that the same linear relationship could explain attainment and age in the two adjacent year groups.

Table 5.5:Regression discontinuity models with primary school data[1,2]

	Fixed effects				**Random effects at school-level**			**Random effect at student-level**
Model	**Intercept**	**Grade level**	**Age**	**Age*Grade level**	**Variance of the intercept**	**Variance of grade level effect**	**Covariance between the grade effect and the intercept**	**Variance of the intercept**
empty model	4.507 (.041)	---	---	---	.204 (.033) *.420 (.068)*	---	---	.204 (.033)
Model 1	4.204 (.044)	.616 (.03) *.635 (.031)*	---	---	.188 (.032) *.388 (.066)*	---	---	.763 (.015)
Model 2	4.446 (.048)	.133 (.047) *.137 (.049)*	.487 (.036) *1.004 (.076)*	---	.187 (.032) *.386 (.066)*	---	---	.745 (.015)
Model 3	4.422 (.047)	.139 (.047) *.144 (.049)*	.484 (.036) *.998 (.074)*	---	.216 (.036) *.446 (.074)*	.04 (.015) *.082 (.032)*	-.04 (.017) *-.082 (.034)*	.736 (.015)
Model 4	4.438 (.052)	.142 (.047) *.147 (.048)*	.517 (.058) *1.066 (.12)*	-.067 (.086) *-.069 (.089)*	.216 (.036) *.39 (.066)*	.04 (.015) *.082 (.032*)	-.04 (.017) *-.082 (.034)*	.735 (.015)

Note. [1]In Model 1 I adjust for the absolute effect of schooling only (denoted as "Grade level" in the table). Model 2 is a random intercept model that makes adjustments for both grade and age. Model 3 is a random slopes models with the effect of grade allowed to vary across schools. Lastly, in Model 4, I include the interaction between age and grade, allowing for the effect of grade to be different in the lower and in the upper grade. [2] Within each cell, the effect size for the corresponding estimate and the associated standard error is given in italics.

5.5.2 Results for Research Hypothesis 3a.4- Research Question 3a.5: The Effect of Adjustments for Student Background Variables on Added-Year Effects/ Investigating Differential School Effectiveness

In this set of statistical analyses I investigated whether adjusting for a set of student-level background characetristics in the RD models (see Table 5.4 for a description the set of these variables) led to changes in the inferences on the magnitude of the absolute effect of schooling obtained with the RD approach. Additionally, I studied whether students from different backgrounds in terms of gender, ethnicity, socioeconomic status and home environment were differentially influenced by schooling – thus addressing the issue of differential school effectiveness.

The results regarding the main effects of the student-level variables controlled for in the RD models and their interactions with the effect of one year of schooling are described in detail in section B.2 of Appendix B. Here, I restrict my discussion only to the findings that relate to the impact of adjustments of such variables on the estimates of the absolute effect of schooling obtained with the RD approach. I only briefly refer to the findings on differential school effectiveness, since this is not a major focus of Study 3 (or, more generally, of my thesis).

5.5.2.1 The effect of adjustments for student background on added-year effects:

In Table 5.6, I display the estimates for the effect of schooling (denoted as "Grade level" in the table) and the effect of age obtained with the RD model that was found to best fit the primary school data (Model 3 in Table 5.5) with ("Adjustments for student-level covariates" in Table 5.6) and without ("Basic model" in Table 5.6) adjustments for the set of student background variables considered in my study.

It can clearly be seen that these estimates are not substantially different when compared across the two models. Therefore, the estimates of the absolute effect of schooling (and the effect of chronological age) obtained with the RD approach are robust to adjustments for student background characteristics – consistent with research hypothesis 3a.4.

5.5.2.2 Interaction of student background with grade: Evidence for differential school effectiveness

The results of this analysis are displayed in Table B.1 in Appendix B (see the column that relates to the analysis with the primary school data). None of the student background variables considered in the analysis moderated the effect of schooling; interestingly enough there is no evidence for differential school effectiveness even in these simple models in which each covariate was considered individually.

Table 5.6: The impact of adjusting for the main effects of significant background variables on the absolute effect of schooling and the effect of chronological age with primary school data.

	Fixed effects			Random effects at school-level			Random effect at student-level
Model	Intercept	Grade level	Age	Variance of the intercept	Variance of grade level effect	Covariance between the grade effect and the intercept	Variance of the intercept
Basic model	4.422 (.047)	.139 (.047) *.144 (.049)*	.484 (.036) *.988 (.074)*	.216 (.036) *.446 (.074)*	.04 (.015) *.082 (.032)*	-.04 (.017) *-.082 (.034)*	.736 (.015)
Adjustments for student-level covariates	4.505 (.082)	.200 (.049) *.206 (.05)*	.378 (.039) *.780 (.08)*	.137 (.029) *.282 (.06)*	.033 (.016) *.068 (.032)*	-.021 (.014) *-.044 (.028)*	.636 (.014)***

Note. The basic model is a random slopes model, with the effect of added-year (denoted as "Grade level" effect in the table) allowed to vary across schools. The effect size for the estimate of the absolute effect of schooling and the effect of age is given in italics.

5.5.3 Results for Research Question 2a.6: Examining Relationships between Schools' Composition and Added-Year Effects

Having fitted the regression discontinuity models in the multilevel modelling framework, the relationship between school-level variables, absolute schooling effects and students' achievement could be explored simultaneously. Between school differences in their added-year effects can be explained in two different ways - in an analogous way to how relative school effects in value added models of educational effectiveness can be explained (see section 2.3.5; see also section 2.3.1 in the literature review): Either by differences in the composition of the student body, that is, by a set of factors over which the school has not control, or by the school's practices and processes (see also section 2.3.4 in the literature review). Here I summarize the results related to the extent to which differences between schools in their absolute effects could be explained by school composition variables. I considered school average achievement (see "School Composition" under section 5.3.2) as this is used to quantify a school's composition.

5.5.3.1 Distinguishing between the effect of school-level variables on random intercepts and random slopes in RD models

A distinction should be made between the effects of school-level variables on random intercepts in regression discontinuity models and on the effects of the same variables on random slopes for the absolute effect of schooling in the RD framework (see equation 2.4 in section 2.4.3 which displays the basic RD model and section B.1 in Appendix B). A significant effect of a school-level variable on random intercepts refers to differences in the mean achievement of students who are in the same year of schooling and have their age at the cut off. On the other hand, a significant effect of a school-level variable on the random effect of the absolute effect of schooling refers to differences in the effect of one extra year of schooling on students' progress, thus on the growth in the average achievement of students at the cut off after receiving an extra year of education. This latter effect is what should be used to make inferences on the effects of

school composition using the regression discontinuity approach. The former is just an indication of the relationship between the school-level variable and students' final outcomes, but without any adjustments for prior achievement and with reference to students at the cut-off point only.

5.5.3.2 The effect of school average attainment on added-year effects

Effect on intercept: The effect of school average achievement on random intercept was positive (β = 011, *se* = .001), suggesting that students with age at the cut off that were in the same year would have higher achievement once they attended a school which consisted of higher achievement students. Note that this effect refers only to the final outcomes of students, not to their individual progress.

Effect on the random slope of the effect of one extra year of schooling: No significant effects of school average achievement were found in explaining the variability of random slopes (β = -.001, *se* = .001) for the effect of one extra year of schooling. Added-year effects were found not to be dependent on the academic achievement of the student body.

Conclusion

The findings for this set of analyses suggest that school composition, as quantified using school average achievement, has no effect on added-year effects – the measure of school effectiveness in regression discontinuity designs.

5.5.4 Results for Research Hypothesis 3a.7-Research Question 3a.8: Integrating Multilevel Structural Equation Models with Regression Discontinuity Designs

The purpose of the analyses presented here was two-fold: First the investigation of the impact of adjustments for measurement error in students' achievement on regression discontinuity estimates– this is used as the response variable in RD designs. Second, the use of multilevel structural equation models to investigate the impact of school composition variables on added-year outcomes after adjustments for measurement error at level 1 and, where relevant, measurement error and sampling error at level 2.

5.5.4.1 The impact of adjustments for measurement error in the students' achievement measures on regression discontinuity estimates

Measurement error reliability for students' mathematics' achievement in primary schools, as given by the ratio of true variance of the latent scores (Empty model/adjustments for measurement error in Table 5.7) divided by the variance of the observed scores (Empty model in Table 5.7) was found to be .91[4]. As expected (see research hypothesis 3a.7; see also section B.3 in Appendix B on the impact of measurement error in achievement scores on RD estimates) given the high reliability of the observed scores, the regression discontinuity estimates and associated standard errors were approximately the same for the manifest and latent achievement multilevel random slopes model. This is also displayed in Table 5.7 in which the RD estimates for Model 3 -- the random slopes regression discontinuity model that was initially fitted to the primary school data (see Table 5.5) -- are displayed beside the corresponding estimates derived

[4] In Table 5.7 the total variance of the observed scores can be calculated as the sum of the school-level intercept variance and the student-level intercept variance corresponding to the "Empty model". In an analogous way, the total variance of the latent scores can be estimated by the sum of the school-level intercept variance and the student-level intercept variance corresponding to the "Empty model/adjustments for measurement error"

from a multilevel structural equation model that made adjustments for measurement in students' achievement through the use of multiple indicators (see "Achievement measures" in section 5.3.2 in Methodology for a description of the manifest variables that were used as multiple indicators for latent mathematics achievement).

5.5.4.2 Investigating the effect of school average achievement in mathematics on added-year effects using multilevel structural equation models

In this analysis, latent school average achievement was not found to moderate significantly the relationship between an extra year of schooling and individual achievement in mathematics: Though non-significant, the effect of the interaction between the absolute effect of schooling and school average achievement was negative ((β = -.045, *se* = .081). This result suggested that the schools' absolute effects on their students were essentially independent of the school average achievement of the students' across the two year groups.

Table 5.7: The impact of adjusting for measurement error in the basic regression discontinuity models with primary school data[1,2]

	Fixed effects			**Random effects at school-level**			**Random effect at student-level**
Model	**Intercept**	**Grade level**	**Age**	**Variance of the intercept**	**Variance of grade level effect**	**Covariance between the grade effect and the intercept**	**Variance of the intercept**
empty model	4.507 (.041)	---	---	.204 (.033) *.420 (.068)*	---	---	.850 (.018)
Empty model/ adjustments for measurement error	---	---	---	.193 (.036) *.398 (.074)*	---	---	.767 (.017)
Model 3	4.422 (.047)	.139 (.047) *.144 (.049)*	.484 (.036) *.998 (.074)*	.216 (.036) *.446 (.074)*	.04 (.015) *.082 (.032)*	-.04 (.017) *-.082 (.034)*	.736 (.015)
Model 3/ adjustments for measurement error	---	.169 (.046) *.183 (.05)*	.430 (.039) *.932 (.084)*	.196 (.035) *.213 (.076)*	.037 (.012) *.08 (.028)*	-.024 (.016) *-.054 (.034)*	.659 (.016)

Note. [1]The Empty model just provides a decomposition of the total variance of the dependent variable in the model (observed mathematics achievement) into a between-school component and a within-school component. When adjustments are made for measurement error (Empty model/adjustments for measurement error) the variability of latent mathematics achievement is decomposed into a within-school and a between-school component. Model 3, the best fitting regression discontinuity model among the four different models tried for primary school data. This is a random slopes models with the effect of grade allowed to vary across schools. [2]Within each cell, the effect size for the corresponding estimate (and associated standard error) is given in italics.

5.5.5 A Summary of the Findings for Study 3a

In summarising the findings for Study 3a, I note the positive and significant effect that one extra year of schooling exerted on students' achievement. Age was also found to have a positive effect on achievement – naturally, since older students are more mature academically. The linear relationship describing age and achievement was the same in the lower (year four) and the higher year group (year five). The regression discontinuity estimates were robust to adjustments for student background characteristics. Adjustments for measurement error in the students' achievement – the criterion in regression discontinuity models also had minimalimpact on the regression discontinuity estimates. School average achievement was used in my models to explain between-school differences in the absolute effect of schooling. A negative but non-significant effect of school average achievement on schools' added-year scores was found. This was the case both when the conventional multilevel modelling was incorporated and when the multilevel structural equation modelling framework was used.

5.6 Results for Study 3b: Secondary School Data (Years Eight and Nine)

5.6.1 Results for Research Hypothesis 3b.1 – Research Hypothesis 3b.3: The Absolute Effect of Schooling and the Effect of Chronological Age on Students' Mathematics Achievement

In the same way as for primary school data, the set of nested models tried for secondary school data is displayed in Table 5.8. The model that fitted the data best was Model 2 (the random intercept regression discontinuity model). The effect of one extra year of schooling was again positive and statistically significant (γ_{20} = .129, *se* = .052, *ES* = .146). The absolute schooling effect did not vary significantly across schools; there did not seem to be significant differences between schools in their absolute effects. The effect of age on achievement was positive and statistically significant (γ_{10} = .173, *se* =.046, *ES* =.394). The relationship between age and achievement was, again found to be the same in the two adjacent grades - no statistically significant interaction was detected between the effect of age and the effect of grade.

Table 5.8: Regression discontinuity models with secondary school data[1,2]

	Fixed effects				**Random effects at school-level**			**Random effect at student-level**
Model	**Intercept**	**Grade level**	**Age**	**Age*Grade level**	**Variance of the intercept**	**Variance of grade level effect**	**Covariance between the grade effect and the intercept**	**Variance of the intercept**
empty model	4.838 (.043)	---	---	---	.215 (.035) *.490 (.078)*	---	---	.574 (.018)
Model 1	4.868 (.047)	.298 (.028) *.339 (.032)*	---	---	.212 (.035) *.482 (.078)*	---	---	.553 (.017)
Model 2	4.772 (.054)	.129 (.052) *.146 (.059)*	.173 (.046) *.394 (.104)*	---	.211 (.034) *.478 (.078)*	---	---	.551 (.017)
Model 3	4.676 (.047)	.131 (.053) *.149 (.06)*	.172 (.046) *.390 (.104)*	---	.233 (.038) *.530 (.086)*	.014 (.013) *.032 (.028)*	-.026 (.014) *-.06 (.032)*	.547 (.017)
Model 4	4.673 (.047)	.161 (.074)* *.183 (.085)**	.204 (.06) *.464 (.136)*	-.059 (.085) *-.067 (.097)*	.232 (.037) *.528 (.084)*	.013 (.012) *.030(.028)*	-.03 (.016) *-.134 (.194)*	.547 (.017)

Note. [1]In Model 1 I adjust for grade. Model 2 is a random intercept model that makes adjustments for both grade and age. Model 3 is a random slopes models with the effect of grade allowed to vary across schools. Lastly, in Model 4, I include the interaction between mage and grade, allowing for the effect of grade to be different in the lower and in the upper grade. [2] Within each cell, the effect size for the corresponding estimate is given in italics.

5.6.2 Results for Research Hypothesis 3b.4- Research Question 3b.5: The Effect of Adjustments for Student Background Variables on Added-Year Effects/ Investigating Differential School Effectiveness

In Table 5.9, I summarise the estimates of the absolute effect of schooling for secondary schools as given by the regression discontinuity approach both before and after adjustments for the main effects of the student background variables considered in my analysis (see section 5.3.2 and the description that I give there on the student background variables considered). The absolute effect of schooling was found to be similar both with (γ_{20} = .129, *se* = .05, ES = .149) and without (γ_{20} = .129, *se* = .052, ES = .146) adjustments for student background. This finding could suggest that the RD estimates of the absolute effect of schooling are robust to adjustments for level 1 covariates of student background. This is, indeed, one of the advantages of the approach over traditional approaches of educational effectiveness which provide estimates of relative school effects and which are very much dependent on the range of student background characteristics controlled for in the models (see section 5.8 in the discussion chapter).

Again, no significant interactions were detected between the student-level background variables considered and the absolute effect of schooling. Note here that, in the same way as for Study 3a, a more elaborate description of the findings regarding the main effects of the student-level covariates considered in my analysis and their interaction with grade-level effects is given in section B.2 of Appendix B.

Table 5.9: The impact of adjusting for the main effects of significant background variables on the absolute effect of schooling and the effect of chronological age with secondary school data

	Fixed effects			Random effects at school-level	Random effect at student-level
Model	Intercept	Grade level	Age	Variance of the intercept	Variance of the intercept
Basic Model	4.772 (.054)	.129 (.052)	.173 (.046)	*.211 (.034)*	.551 (.017)
		.14 (.059)	*.394 (.104)*	*.478 (.078)*	
Adjustments for student-level covariates	5.646 (.168)	.129 (.05)	.157 (.045)	*.131 (.024)*	.480 (.014)
		.149 (.027)	*.356 (.102)*	*.294 (.114)*	

Note. The basic model is a random intercept model. The effect size for the estimate of the absolute effect of schooling and the effect of age are given in italics.

5.6.3 Results for Research Question 3b.7: Examining Relationships among Schools' Composition and Added-Year Effects

Although the effects of variables relevant to the schools' composition on the schools' added-year effects could still be tested and reported with secondary school data, I decided not to proceed with this analysis: the practical importance of the findings would be minimal since no between school variability in the schools' absolute effects was detected (see section 5.6.1). Besides, I was interested mainly in the effects of school composition with primary school data – in accordance with the aims of the first two studies of my thesis that were also concerned with early primary school data (see section 3.3.3 on the measurement and data samples used in Study 1a; section 3.6.1 on the data used in Study 1b and section 4.4.1 for a description for the data samples used in Study 2).

5.6.4 Results for Research Hypothesis 3b.8-Research Question 3b.9: Integrating Multilevel Structural Equation Models with Regression Discontinuity Designs

The results for the basic RD model with secondary school data before and after adjustments for measurement error in students' achievement are presented here (Table 5.10). In the same way as for primary schools (see "The impact of adjustments for measurement error in the students' achievement measures on regression discontinuity estimates" under section 5.5.4 for primary school data), measurement error reliability for secondary school data was found to be .909, again relatively high. The absolute effect of schooling and associated standard errors were again almost unaffected by measurement error adjustments in the criterion.

Table 5.10: The impact of adjusting for measurement error in the basic regression discontinuity models with secondary school data[1,2]

	Fixed effects			Random effects at school-level	Random effect at student-level
Model	Intercept	Grade level	Age	Variance of the intercept	Variance of the intercept
empty model	4.838 (.043)	---	---	.215 (.035) *.490 (.078)*	.574 (.018)
empty model/ adjustments for measurement error	---	---	---	.199 (.033) *.452 (.076)*	.518 (.018)
Model 2	4.772 (.054)	.129 (.052) *.146 (.059)*	.173 (.046) *.394 (.104)*	*.211 (.034)* *.478 (.078)*	.551 (.017)
Model 2/ adjustments for measurement error	---	.120 (.049) *.168 (.068)*	.154 (.043) *.430 (.12)*	.211 (.036) *.588 (.100)*	.486 (.018)

Note. [1]Model 2 is the best fitting regression discontinuity model among the four different models tried for secondary school data; this is a random intercept model. [2]Within each cell, the effect size for the corresponding estimate is given in italics.

5.6.5 Results for Research Hypothesis 3b.9: The Estimated Effects of Schooling and Age on Attainment are Larger for Primary School as Compared to Secondary School Data.

My focus here is on comparing the estimates of the basic RD models across primary school and secondary school data

Absolute effect of schooling

The effect of schooling was found to be positive both for primary (γ_{20} = .139, *se* = .047, ES = .144) and secondary school data (γ_{20} = .129, *se* = .052, ES = .146); its magnitude was somewhat smaller for secondary schools as compared to primary schools. This implies that, according to the data analysis presented, schooling does exert a (positive) impact on students' cognitive progress and it seems to be important both for primary and secondary school data.

Effect of chronological age

The effect of chronological age was also positive and significant for primary (γ_{10} = .484, *se* = .036, ES = .998) and secondary school (γ_{10} = .173, *se* = .046, ES = .394). The within school coefficient for age was the same in the lower grade as in the higher grade for both populations; this was much larger for students in their earlier years of schooling (years four and five) as opposed to the students in their later years (years eight and nine); in accordance to my research hypotheses.

5.6.6 Relative Effects of Schooling: The Variance of the Absolute Effect of Schooling across Schools.

Despite the fact that significant between school variance was found for the absolute effect of schooling in primary schools ($Var(U_{1j})$ = .04, se = .015, ES = .082), evidence for relative school effects across schools - no variation was detected for secondary school data.

5.6.7 A Summary of the Findings for Study 3b

In Study 3b, concerning English secondary school data (years eight and nine), a positive effect of one extra year of schooling on students' achievement was detected; the differences across schools in these positive absolute effects were not found to be statistically significant. The effect on one chronological year was also found to be positive and statistically significant and the linear relationship describing age and achievement was the same in the lower (year eight) as well as in the higher (year nine) year group. The regression discontinuity estimates in this analysis, just like those concerning primary school data, were not changed substantially after adjustments for the main effects of student background variables or for measurement error in individual achievement – the criterion in regression discontinuity models.

Chapter 6: Discussion/Conclusion

Complex educational issues require sophisticated methodologies to ensure that invalid conclusions are not to be drawn. All three studies of my thesis are methodological-substantive synergies which use cutting-edge methodology to address a highly debated topic in educational research: the potential effect of school composition, in terms of school average achievement, on students' cognitive outcomes (mathematics achievement; see Study 1 and Study 3) and affective outcomes (mathematics self-concept; see Study 2).

In the first study of my thesis (Study 1) I consider the extent to which school-level aggregates of achievement explain relative school effects when adjusted for in value added models of educational effectiveness (compositional analysis models that control for prior achievement). In the third study (Study 3), I use the same school-level variable (school average mathematics achievement) to explain between school differences in their added-year scores – an alternative measure of school effectiveness to conventional value added measures.

Studying the impact of educational institutions on students' cognitive outcomes is just as important as evaluating their influence on the students' affective outcomes: In Study 2, I investigate the compositional effect of school average mathematics achievement on students' self-concept with respect to mathematics. In this respect, I seek to verify a well-established phenomenon in educational psychology, the Big-Fish-Little-Pond-Effect (BFLPE) hypothesis.

The question of whether or not school composition variables, such as the academic excellence of the school's intake, can actually have an impact on students' individual outcomes is an issue with important theoretical and substantive implications. It is relevant to the question whether schools do indeed "matter" and can make a difference to students' achievement, compensating for pre-existing differences in achievement due, for example, to social

inequalities. It is also related to the impact of practices such as ability grouping and selective schooling on students' progress.

The "traditional" approach to measuring compositional effects is inherently biased under most conditions and is suboptimal. This has been clearly demonstrated in relation to simulated data (Lüdtke, Marsh, Robitzsch, Trautwein, Asparouhov and Muthén, 2008; Lüdtke, Marsh, Robitzsch, and Trautwein, 2011) and real data (Marsh, Lüdtke et al., 2012; Marsh, Lüdtke, et al., 2009). Similarly, I note that the critical Harker and Tymms (2004) study identified the "phantom effect" associated with measurement error that is inherent in the "traditional" approach, but offered no solutions to the problem. An important contribution of my study is to demonstrate the application of stronger models of compositional effects to the school effectiveness literature, models that are well established in the broader area of research on compositional effects.

The Marsh, Lüdtke et al. (2009; see also Lüdtke et al., 2011) methodological framework used throughout my thesis (see section 1.4 and section 2.2.5) extends the classical test theory framework to address simultaneously random error due to the sampling of items (measurement error) and random error due to the sampling of people (sampling error). The models apply corrections for (a) bias due to measurement error in individual variables and the corresponding aggregates and (b) bias due to sampling error in the aggregation of individual-level variables to form the school-level aggregates. Hopefully, by employing these classes of models in their research, educational researchers will be able to test hypotheses based on a more reliable methodological approach.

6.1 The Marsh, Lüdtke et al. 2x2 Taxonomy of Models and other Approaches that can Control for Measurement and/or Sampling Error Bias in Compositional Analysis

The issue of inaccuracies due to measurement error and the potential consequences that this can have for the validity of the inferences in educational research is not new. In fact, an extensive literature has been developed which is devoted explicitly to the derivation and accuracy of test scores (Marcoulides and Kyriakides, 2010); this is the so-called psychometric literature. The reliability of aggregate variables has also been discussed in other methodological contexts (e.g., within the generalizability theory literature; see Cronbach, Linn, Brennan and Haertel, 1997). In the following section I explain the way in which the Marsh, Lüdtke et al. (2009) framework complements other approaches which can also deal with measurement and sampling error in statistical analyses.

6.1.1 Simultaneous Adjustments for Measurement Error, Sampling Error and for Hierarchical Structures in the Underlying Data

Methods that accommodate measurement error in single – level analysis are well known in the educational community (Degracie and Fuller, 1972; Fuller, 1987; Joreskog, 1970; Plewis, 1985). These have been extended to multilevel analyses by researchers such as Woodhouse et al., (1996) and Browne, Goldstein and Rasbash (2001). The approach by Woodhouse et al. (1996) is based upon moment type estimators while Browne at el., (2001) developed an algorithm using Markov Chain Monte Carlo (MCMC) estimation which they implement into the MLwiN software (Browne, 2004). More recent advances have been proposed by Ferrão and Goldstein (2009) and Goldstein, Kounali and Robinson (2008). These propose a two-stage approach to control for measurement error. Information on the reliability of the level-1 variables is obtained externally and then a range of assumed values can be incorporated into the estimation process via an MCMC estimation algorithm, "in the spirit of sensitivity analysis"

(Goldstein et al., 2008). The advantage of the 2x2 taxonomy of compositional models (Marsh, Lütdke et al., 2009) over previous attempts to control for measurement and sampling error in the educational data (see section 2.2.5 in the literature review) is that they account not only for measurement error at level 1 and level 2 but also for sampling error (see the section 2.2.5). Of course, Woodhouse et al. (1996) also considered sampling error but the formula that they proposed dealt only with finite clusters (see further comments on distinguishing between a formative and a latent aggregation process" in section 6.2). Hutchinson also discussed measurement error in compositional models in combination with sampling error arising when the set of pupils involved in the analysis represent only "a random sample from a large population of possible pupils" (Hutchinson, 2007, p. 222). Still, the Marsh, Lütdke et al., (2009) framework has several advantages as opposed to these alternative approaches that allow adjustments for measurement and sampling error, especially when a researcher deals with large datasets and is much more flexible in terms specifying alternative model of measurement error in more complex models (e.g., method effects or correlated uniquenesses when the same items are used on multiple occasions).It can relatively easily be implemented within the Mplus statistical package (Muthén and Muthén, 1998 - 2012). Moreover, the four basic models of the 2x2 taxonomy can be extended to more complex models, for instance, where an explanatory variable that contains measurement error has a random coefficient. Lastly, the Marsh, Lütdke et al., (2009) framework does not require any external assumptions about the reliability of the data; this is estimated as part of the modelling procedure.

6.1.2 Generalizability Theory: An Alternative Methodological Framework in which the Reliability of the Aggregated Variables is addressed

When assessing the reliability of aggregate variables, it is crucial to simultaneously address reliability in relation to measurement error (due to sampling of items) and sampling error (due to sampling of people). Generalizability theory (e.g. Cronbach, Gleser, Nanda and Rajaratnam, 1972; Cronbach, Linn, Brennan and Haertel, 1997) provides an alternative framework to that

adopted in my study, namely multilevel structural equation modelling, which can be used to address random error due to the sampling of items (measurement error) and due to the sampling of people (sampling error) in group-level constructs. This is also a more recent and more liberalized approach to measurement than the Classical Test Theory (see Creemers, Kyriakides and Sammons, 2010, chapter 10) , the methodological basis for the Marsh, Lüdtke et al. (2009) 2x2 taxonomy of models.

In the generalizability theory framework (Brennan, 2001; 2010), an investigator must define the potential sources of error; each of these is referred to as a facet. Then the purpose is to determine the extent to which scores are comparable across the different conditions underlying a facet. Suppose, for example, that the same test is administered to the same group of people on two different occasions. This is a single facet study and the dimension in question is occasion. When this facet explains a significant amount of the variance in the observed scores, then the findings do not generalise across levels (the two different times of administrating the test) of that facet. In this case such studies, referred to as generalizability or G-Studies, yield a poor "generalizability coefficient" and the facet is examined further. Ideally, a G-study should provide evidence that one can generalize across levels without misinterpreting the data. With generalizability theory it is possible to dis-confound the multiple sources of error that can be of interest in an even more abstract manner (for instance errors attributable to different testing sessions, different test items or different modes of administrating a test).

Although the approach can effectively be used to address psychometric properties of latent constructs and can be used to provide in the standard error a suitable indicator for uncertainty (e.g. Cronbach et al., 1997), it does not provide a methodological framework on how to investigate relationships between constructs after potential sources of error are taken into account. This can be regarded as a limitation of the application of generalizability theory against the models we incorporate for the purpose of our article.

6.2 Formative and Reflective Aggregates: Correcting for Sampling Error Using Doubly Latent Approaches

The partial (manifest latent and latent manifest) and full correction (doubly latent) compositional models (see section 2.2.5 in the literature review and Table 2.1 regarding the Marsh, Lüdtke et al. 2009 taxonomy of model) should be applied with appropriate caution. The doubly latent model is not a panacea; on to the contrary, under certain circumstances the partial correction approach or even the doubly manifest approach should be used for the investigation of compositional effects.

Whenever applied researchers use the Marsh, Lüdtke et al. (2009) 2x2 taxonomy of models to evaluate compositional effects, they have to face the dilemma of which of the four estimates obtained by each approach would best suit their data. While measurement error adjustments should always yield a compositional effect estimate free of measurement error bias, sampling error adjustments are not always justifiable. For instance, the type of aggregate variable that a researcher deals with plays a crucial role in determining the approach that should be applied in the analysis in order to obtain appropriate compositional effect estimates.

In the literature, two types of aggregate variables can be distinguished, namely "formative" and "reflective" aggregates. The key distinction between formative and reflective variables is the referent in the level 1 measure. For formative constructs, the focus of the level 1 indicator is on individual-level characteristics. The level 2 construct is caused by the lower-level observations and describes the group composition. Thus, for example school average achievement, the construct that I use in my thesis, is a classic example of a formative construct (but see also section 6.3). Other examples of formative aggregates include level 2 aggregations of gender, socioeconomic status or any other background or demographic characteristics of individuals within a group.

For reflective constructs, on the other hand, the purpose of the set of level 1 measures is to directly assess a group-level characteristics. Individuals rate some aspect of the group rather than an individual characteristic regarding their own characteristics. Examples of reflective constructs include aggregates of student-level ratings of the class climate or on the extent to which a teacher is effective or organised.

Distinguishing between formative and reflective aggregates is closely related to the sampling process underlying the construction of the aggregated construct. For formative constructs, a finite (formative or manifest) aggregation process is assumed. This implies that the assumed population of the group is finite – represented by the individual-level units of which the group is comprised. On the other hand, for reflective constructs a latent (reflective or infinite) sampling process is assumed. The value of the aggregated construct is just one manifestation of all the potential values that it could have taken using finite subsets of the infinite population of the units that form the group.

The latent approaches from the 2x2 taxonomy (the manifest latent approach and the doubly latent approach) correct for sampling error assuming a latent aggregation process. Although this is justifiable when reflective constructs are involved in the analysis, more careful consideration should be used in deciding whether or not this is appropriate when formative constructs are used in the analysis.

When the focus is on formative constructs, educational researchers should consider sampling ratio, that is, the proportion of level 1 individuals sampled from within each group in their analysis. When the sampling ratio approaches one, that is, almost all the lower-level units from the higher-level group are involved in their analysis, educational researchers should use a manifest aggregation approach rather than a latent aggregation approach: in this case, the use of doubly latent model may potentially overcorrect for sampling error in the compositional effect estimate and thus lead to positively biased compositional effects. On the other hand, when the sampling ratio is small, then educational researchers should consider using latent aggregation

approaches (i.e. the doubly latent model). Indeed, in a simulation study, Lüdtke et al. (2008) showed with simulations that for a sampling ratio as small as 20 per cent, latent aggregation approaches outperform the manifest aggregation approach both in terms of bias and variability in the assessment of the effects of formative constructs. Still, the researchers recommend the comparison of both manifest and latent aggregation models to determine if the difference is substantively meaningful; logically the best estimate lies somewhere between these two estimates.

Generally, the recommendation (see Marsh, Lüdtke et al., 2009) is to juxtapose all the four different approaches proposed in the 2x2 taxonomy. Educational researchers can then observe what difference it makes to the estimates when adjusting for different sources of error. Ideally all the estimates should be close to each other—if all constructs have small amounts of measurement error and there are many level 1 units (i.e., students) in each level 2 group (i.e., schools).

6.3 School Average Achievement: A Formative or Reflective Construct?

As I have already explained (see previous section), school average achievement is typically assumed to be a formative construct. What is interesting is that under certain circumstances, it can be approximated by a reflective aggregation process. When, for instance, a small proportion of level 1 units are sampled from each level 2 unit resulting in a rather small sampling ratio, then the sampling "mimics" reflective aggregation where only a finite sample is used from an infinite population.

Goldstein, Kounali and Robinson (2008; see also Shin and Raudenbush, 2010) give an example from educational data in which the school average achievement can be assumed to be a proxy for that school's long term characteristics, and, thus, based on a reflective aggregation process even if it is in fact a formative construct. The observed school average of a specific year may be viewed as a realization of all the possible values that this school-level construct can take

over all the academic years. In the literature (Särndal, Swensson and Wretman, 2003) a relevant idea refers to a super-population from which the observed data are generated, in such a way that even when all the units are sampled from within a cluster, the properties of the population are still not known precisely.

Authors such as Burstein (1980) are hesitant to accept such claims. They express the view that a pupil's progress is more likely to be influenced by the pupils who are his or her actual classmates and not by the totality of pupils that there might have been. This is a view I also share – school average achievement should most reasonably be claimed to be a formative construct.

6.4 Study 1: A Discussion of the Findings

6.4.1 Correcting for Positive Measurement Error Bias in Compositional Effects Estimates

Study 1, based on English data, revealed a weak but significant negative compositional effect of school average ability. When adjustments were made for measurement error, this negative compositional effect became more negative, with effect size estimates almost doubling in magnitude. Study 2, based on Cypriot data, initially resulted in a small but significantly positive compositional effect. This is in line with previous research conducted using early primary school mathematics achievement data from Cyprus (e.g., Kyriakides, 2008; Kyriakides and Creemers, 2008). The effect disappeared – it became non-significant -- after adjustments for unreliability due to measurement error. The sign of the compositional effect was even reversed, it became negative. One reason why no significant compositional effect was detected in the analysis with the Cyprus data is that there were no substantial differences in the student intake in the schools in Cyprus in the first place (see also the "Results" section for Study 1b, that is, section 3.7). Thus, the power of the statistical analysis conducted as part of Study 1b was much lower compared to the analysis of Study 1a.

Both studies were consistent in showing that conventional approaches led to positively biased estimates of compositional effects, due to ignoring the problem of measurement error in multilevel modelling. Previous studies therefore are likely to have provided positively biased effects of school composition on students' outcomes.

6.4.2 The Impact of Sampling Error Adjustments on Compositional Effects Estimates

Adjustments for sampling error in the analysis gave larger effects whether these were positive or negative: whenever a positive compositional effect was detected, this became more positive when sampling error was adjusted for; whenever a negative compositional effect was detected, this became more negative once sampling error was controlled for. This is in line with previous research (Lüdtke et a., 2008; Marsh, Lüdtke et al., 2009).

Indeed, mathematical formulae that quantify the bias in the estimated compositional effect when unreliability due to sampling error is not taken into account (see relevant formula in the Appendix) show that, to the extent that sampling error exists in the data, compositional effects are likely to be underestimated. In verifying these mathematical derivations with the English PIPS database (see formula A.2.7 in section A.2.3 of Appendix A, quantifying sampling error bias), I estimated the bias in the compositional effect due to sampling error to be equal to .0058 - rather small. This was consistent with the fact that when adjustments for sampling error were made in my data, the compositional effect remained almost unchanged. When applying the same formula with the data from Cyprus (Study 1b) I estimated sampling error bias in the compositional effect – assuming latent aggregation - to be -.053. The negative sign in the estimate of the bias occurs due to the fact that the compositional effect with the Cyprus data is originally estimated to be positive.

Another interesting observation with the findings of both analyses comprising Study 1 (that concerned with the English data Study 1a, and that concerned with the Cyprus data, Study 1b) is the fact that the impact of measurement error adjustments on the estimates of compositional effects is more dramatic than the impact of controlling for sampling error. This is indicative of the fact that the prevalence of measurement error in the data is more problematic than the prevalence of sampling error. This is likely to be idiosyncratic to this study in which the sampling ratios and number of students/school are relatively large. If the sampling ratios and numbers of students/school had been smaller, the correction for sampling error would have been larger. Also, if there had been no measurement error, correction for measurement error would make no difference.

6.4.3 Differences in Estimates of Compositional Effects of Prior Achievement Obtained by the Application of the Four Models of the 2x2 Taxonomy

In estimating compositional effects in Study 1, I used all four models of the 2x2 taxonomy and thus obtained four different estimates of the effect of interest. In comparing the compositional effects across the four models, the question arises as to which would be the most appropriate approach to estimating school-level effects of achievement in my analyses. What might be problematic is that the manifest latent and the doubly latent approaches could overestimate the variability due to sampling error (see sections 6.2 and 6.3 in the discussion on choosing the appropriate approach from the 2x2 taxonomy). Corrections for sampling error by the two approaches are made on the assumption of a latent aggregation process: The school-level characteristics are conceived as latent unobserved constructs that can be inferred on the basis of a finite subset of a potentially infinite number of observers. Nevertheless, when considering the compositional effect of achievement, it is more reasonable to assume a finite aggregation process: Adjustments for sampling error should be made on the basis that aggregation involves only a subsample of the finite number of units within the group.

Because of this conceptual difference between the aggregation process assumed by the models of the 2 x 2 taxonomy that control for sampling error and that which was actually followed, the latent manifest approach is more likely to result in slightly underestimated estimates of compositional effects for formative constructs, while the doubly latent approach is more likely to result in slightly overestimated estimates. Marsh, Lüdtke, et al. (2009) suggest that the correct estimate lies between that of the latent and that of the doubly latent manifest. What is important in my analysis is that the direction and the size of these two estimates do not differ substantially.

6.4.4 Verification of the Negative Compositional Effect, as this was Detected with the English Primary School Data in Study 1

The negative compositional effect of school average achievement on students' progress in mathematics is intuitively counteractive. Theoretical models seeking to explain the way in which the student body affects academic outcomes (Thrupp, 1999; Alexander, Fennesey, McDill and D' Amico, 1979) typically take it for granted that students in selective schools have an advantage compared to their peers studying in lower achievement institutions. What is even more interesting is that this effect was obtained by just applying the basic compositional analysis model that makes no adjustments for measurement error or other background variables and is therefore likely to lead to biased compositional effects. In order to verify the findings on this negative effect, a number of supplementary analyses were conducted (see section A.4 in Appendix A). These involved replicating the analysis procedure with datasets derived from the original data but using (i) different missing data procedures and (ii) different criteria for the inclusion of schools and students in the sample for the analysis – the datasets derived in the second set of supplementary analysis were all sub-samples of the original data. The negative compositional effect was robust for all the different analyses performed.

6.4.5 Measurement Error and Random Effects Estimates in Compositional Analysis: Methodological and Substantive Implications for the Assessment of the Magnitude of Relative School Effects

The simulation that I performed as part of Study 1a (see section 3.4.6 on the "Results for Research Hypothesis 1a.9 – Research Question 1a.11: The Impact of Measurement Error on Random Effects Estimates"; see also Table 3.4) revealed that in multilevel compositional models which control for prior (mathematics) achievement and average prior (mathematics) achievement, measurement error in individual-level prior achievement results in a larger estimate of the within-group variance in a compositional model. The between-group variance remains almost unaffected by measurement error at level 1.

Random effects in compositional models are of special substantive value for value added analysis of educational institutions (see section "The use of compositional effects in value added models of educational effectiveness" in section 2.3.5; see also section A.1 in Appendix A for a more technical presentation of the use of compositional effects in value added models). The level 2 residuals that are obtained after adjustments for prior achievement and average prior achievement in compositional models are often used in educational effectiveness research to obtain the value added scores of educational institutions. Then school effects are typically obtained by dividing the level 2 residual variance with the total variability in the students' final achievement (the sum of the within-group variance and between-group variance in the student's final achievement). Note here the difference between the term "school effects" and the "school *compositional* effects" in the way these are used in my thesis: The former relates to the random part of value added models while the latter relates to the fixed effect of a school-level aggregate over and above the effect of the corresponding variable at the individual-level. The impact of measurement and sampling error adjustments on compositional effects estimates in Study 1 have been discussed in sections 6.4.1 ("Correcting for Positive Measurement Error Bias in Compositional Effects Estimates") and 6.4.2 ("The Impact of Sampling Error Adjustments on

Compositional Effects Estimates"). Here I discuss the impact of measurement error adjustments on random effects estimates in compositional models and on the derived school effects estimates obtained with value added models of educational effectiveness.

My analysis (see Table 3.4) showed that, in the multilevel modelling framework where no adjustments for measurement error are made, measurement error unreliability in prior achievement scores results in smaller school effects estimates. This follows naturally from the fact that measurement error at level 1 results in a larger estimate of the within-group variance (that is used in the denominator of the ratio that is used to estimate school effects; see section A.1.3 in Appendix A.1 on the exact formula used to estimate school effects in compositional models used in the value added modelling framework) while the impact on level-2 variance (the numerator in the ratio used to estimate school effects) is not so severe.

What is of special importance in my analysis is that the use of the latent-manifest and doubly latent approaches effectively corrected for measurement error in prior achievement (but also in subsequent achievement) in compositional models, resulting in a smaller estimate of the level 1 residual variance (see "Measurement error bias in random effects" under section 2.2.3). In this way, the school effects estimates obtained after corrections for measurement error (see Table 3.4) were larger in magnitude.

Methodologically, my results are in line with those of Woodhouse, Yang, Goldstein, and Rasbash (1996) who showed that adjustments for measurement error in prior achievement leads to an increase in the relative size of the overall school effect. A range of studies have also confirmed this suggestion (e.g. Ferrão and Goldstein, 2008; Guldemond and Bosker, 2009; Goldstein, Kounali, and Robinson; 2008). Importantly, with the value added models that I used in my analysis where I adjusted for average prior achievement as well as for individual prior achievement, the differences in the estimates of school effects obtained before and after adjustments for measurement error were trivial. This is in line with Fletcher (2012) who, in a relevant study that he conducted, suggested that the 2x2 taxonomy of models did not have a

much effect on school-level residuals(and, accordingly on school effects estimates), as long as the compositional effects were retained (see also a similar claim in "Measurement error bias in random effects" under section 2.2.3).

From a substantive point of view, my findings are relevant to the on-going discussion (Gorard, 2010; Muijs, Kelly, Sammons, Reynolds and Chapman, 2012; Reynolds, Chapman, Kelly, Muijs and Sammons, 2011) that began with Gorard (2010) who claimed that the problem of measurement error makes the estimates of school effects obtained with the use of multilevel models invalid. Gorard argued that the impact of measurement error on school-level and particularly on individual-level residual variance in value added models of educational effectiveness results in misleading estimates of school effects indicators: he claims that school effects are the result of "propagate" errors (Gorard, 2010, p.753) and that, in reality, school effects are too small to be regarded as of any substantive value. Obviously, the findings of the present study contradict these erroneous inferences: measurement error actually results in smaller school effects estimates and when it is adjusted for, school effects become larger. In this way, school effects estimates reported in current studies of educational effectiveness that do not adjust for measurement are likely to be underestimated rather than overestimated – this is in contrast with Gorard's suggestions and in line with the claims of Muijs et al. (2011; see also Reynolds et al., 2012).

6.4.6 Directions for Further Research

The methodological framework used in Study 1 to investigate school compositional effects can be used to address a broader range of issues that are of substantive importance in the EER paradigm. This last part of my discussion on the findings of Study 1 is dedicated to outlining some of these. On one hand, this enables the reader to place the focus of enquiry of my thesis in the wider research paradigm. On the other hand, it serves to delineate the way in which the present research can be extended in future research.

Teacher effects

The present thesis can be characterised as a school effectiveness study rather than a teacher effectiveness study (see section 2.3.5 on the definition of "educational effectiveness research" and "school effectiveness research"), since it considers the effect of school composition on students' outcomes and the assessment of the effect of one extra year of schooling (see Study 3). Nevertheless, a view which has gained popularity among educational researchers (Kyriakides, Creemers, Teddlie and Muijs, 2010; Marsh, Nagengast, Fletcher and Televantou, 2011; Tymms, Jones, Albone and Henderson, 2009) is that the impacts of classrooms or teachers are much more significant than the effects of schools in explaining variation in achievement. Within each school, different teachers may impact on student achievement in different ways. Factors such as teacher behaviour, teacher expectations and classroom organization may result in within school variability in students' performance between teachers. This may lead to substantially larger teacher effects than school effects (Marsh et al., 2011).

The importance of teacher effects has been recognized in the United States, where much of the research on school effectiveness concerns the effects of individual teachers (e.g. Ballou, Sanders and Wright, 2004; Harris and McCaffrey, 2010). In the UK the focus has been mainly on the assessment of school quality (Marsh et al., 2011). However, recently the Education Select Committee in the UK has proposed a pay system that will reward those teachers adding the greatest value to pupil performance (http://www.publications.parliament.uk/pa/cm201012/cmselect/cmeduc/1515/151505.htm). A relevant study, quantitative in nature - one of the few quantitative studies that have been conducted in UK to investigate teacher effects - reported "considerable variability in teacher effectiveness, a little higher than the estimates found in (…) US studies" (Slater, Davies and Burger, 2011, p.1).

Although the present thesis applies value added models focusing mainly on the estimation of school compositional effects, the models proposed are also relevant to teacher

effectiveness studies. The doubly latent models in particular (see Table 2.1) have the potential to make a substantial contribution to teacher effectiveness research. I note that value added models of teacher effects are likely to be considerably more complex – and, perhaps, problematic – due to complications in controlling for pre-existing differences, the inconsistency of estimates over time and different classes, and disentangling the effects of the many different teachers to whom any given student is exposed (e.g., Marsh et al., 2011).

Stability of school effects

The stability of school effects across time is another issue of concern for school effectiveness research. Stability relates to the extent to which school effects implied by different age groups are similar to each other (Thomas, Sammons, Mortimore and Smees, 1997). There is no clear consensus on whether and to what extent schools vary in their effects, with some studies (e.g.Goldstein, 1987; Rutter, Maughan, Mortimore, Ouston and Smith, 1979) providing evidence of high correlations in the estimated school effects from year to year while others report lack of stability in school effects over time (e.g. Leckie and Goldstein, 2009). The recommendation is that estimates of school effectiveness should be based on data from several years (Sammons and Bakkum, 2011; Sammons, Thomas, Mortimore and Smees, 1997). Although my thesis relates to specific student groups, the same issues could be addressed with data from a different group. This could be the focus of future research, especially bearing in mind the fact that the Performance Indicators at Primary School Project, one of the main databases used for the purpose of my thesis, allows for this potential to be explored.

Persistence of school effects

Several studies within SER (Goldstein, Burgess and McConnell, 2007) address the issue of the continuation of the effects of educational institutions previously attended on students' achievement, even after the student has left the school and entered a new school. For example, when investigating the effectiveness of a secondary school, a researcher could take into account

the impact of the primary school attended on the students' academic outcomes (see Sammons, Nuttall, Cuttance, and Thomas, 1995; Goldstein and Sammons, 1997). To address the issue of continuing primary school effects, cross-classified models can be incorporated (see for example Goldstein, 1995, chapter 8). These extend the traditional multilevel models (see section 2.2.1, presenting the concentional multilevel compositional analysis model) by incorporating into the random part of the models a residual corresponding to the primary school of the students.

In the same way in which a cross classified model is used to control for the simultaneous effects of both the primary school and the secondary school, multiple membership models (Fielding and Goldstein, 2006) can be used to take into account the effect of all the previous educational institutions that an individual may have attended (see Leckie, 2009; Goldstein et al., 2007). Moreover, more than two crossed effects can be incorporated in cross classified models, modelling, for example, the effect of neighbourhood (Leckie, 2009) or Local Educational Authority (Tymms, Merrell, Heron, Jones, Albone and Henderson, 2008) as well as the effect of primary and secondary school on students' attainment.

The models demonstrated in the present thesis can be extended to take into account these complex data structures – although the derived models will be much more demanding computationally.

6.5 Study 2: A Discussion of the Findings

6.5.1 A Summary of the Main Findings of Study 2

The first part of Study 2 (Study 2a) verifies the BFLPE with year one (five-year-old) and year four (seven-year old) students. To be precise, I found weak, negative and a significant compositional effects of school average achievement at the end of year one on (i) students' mathematics self-concept at the end of year one and (ii) students' mathematics self-concept in year four.What makes the findings of Study 2 especially important is that they are derived using

cutting-edge methodological advances in the field of self-concept and BFLPE research, able to accommodate for measurement error in individual- and school-level achievement and sampling error in the aggregation of the lower-level units to form the higher-level aggregates.

In a supplementary analysis to Study 1a (Study 2b) I demonstrate that the negative compositional effect of school average achievement at the end of year one on self-concept at year four – detected in Study 2a – persists even after adjustments for individual self-concept at year one in the model. Study 2b also complements Study 1a, in that it evaluates the extent to which the negative compositional effect of school average achievement at the end of year one on subsequent achievement in mathematics at year four can be explained by the prevalence of a BFLPE in the data underlying the two analyses. Here, I remind the reader that the dataset used for the purposes of Study 2 was the same as that used in Study 1a, namely year one and year four mathematics achievement from the Performance Indicators at Primary School Project (PIPS). Specifically, using mediation analysis, a significant indirect effect (see Table 4.4) of school average achievement in year one on students' mathematics achievement in year four, via year four self-concept was detected.

6.5.2 Verification of the Big-Fish-Little-Pond-Effect in Early Primary Years

With the Study 2 I contribute to existing knowledge in demonstrating the prevalence of a BFPLE for students as young as five to nine years old. Particularly for children in this range, whose reports of self-concept can be highly unreliable, the application of multilevel structural equation models for assessing the magnitude of compositional effects can be especially useful – in this way bias due to measurement error unreliability can be corrected for. Although the BFLPE hypothesis has been widely investigated around the world, there exist surprisingly few UK studies addressing this phenomenon. This is apparently the first study to investigate the phenomenon using multilevel structural equation models for students this young. Some evidence for the prevalence of the BFLPE in early education in England has been suggested by

studies such as that of Tymms (2001). However, Tymms used conventional multilevel modelling approaches with manifest variables and manifest aggregation. He found evidence consistent with the BFLPE hypothesis for mathematics and reading on a sample of Year two students adopted from the same data base as the one used in the present study, namely the Performance Indicators in Primary School (PIPS) project. However, this study did not apply doubly-latent models, understandably given that this was not a major focus of their research and was conducted prior to the time when these models were readily available. In a recent study, Nagengast and Marsh (2011) extended multi-group, doubly-latent, multilevel, structural equation models and used PISA 2006 school data, providing evidence on the BFLPE for the four UK countries (Northern Ireland, England, Scotland and Wales) individually as well as the total UK sample. Nevertheless, their analysis was focused on secondary school data and on science self-concept.

6.5.3 Evidence on the Stability of the BFLPE over Time

What is interesting with the BFLPE models that I evaluate in Study 2 is that the magnitude of the negative compositional effect of school average achievement in year one on students' self-concept in year one is relatively small and only marginally significant. By year four, it grows much larger – it becomes almost seven times as large -- and becomes highly significant.One reason for this could be the fact that social comparison processes and the relation between academic self-concept and achievement are weak in year one. This is different from students over a similar period of time in high school. This is also in line with a suggestion that has been made in previous literature (see for instance, Marsh, Köller and Beaumert, 2001), namely that the longer a student remains in a selective school setting, the larger BFLPEs become.

Indeed, testing the mediation hypothesis whether BFLPEs in year four persisted even after controlling for year one self-concept (see Table 4.3), I found a negative and significant direct effect of school average achievement in year one on students' self-concept in year four.

This implied that the BFLPE was significantly larger for year four as compared to year one students. Negative effects of school average achievement on subsequent achievement even after controlling for earlier BFLPEs have also been reported by Marsh, Kong and Hau (2003). In my study I extend their work (see also similar work by Marsh and O' Mara, 2010) in that (i) I consider students of a much younger age – they used a large sample of secondary school students in Hong-Kong, (ii) I use structural equation models and the "MODEL CONSTRAINT" in Mplus. In this way, in addition to taking into account the multilevel structure in the data, I control for measurement error in individual- and school-level achievement.

6.5.4 An Attempt to Explain the Negative Compositional Effects of School Average Achievement on Subsequent Mathematics achievement

The negative compositional effect of school average achievement on subsequent mathematics achievement found in Study 1a (see section 3.4.7 for a summary of the findings of Study 1a) was in line with Study 2, revealing the prevalence of BFLPEs (negative compositional effect of school average mathematics achievement in year one on students' self-concept in year four) with the same set of data. Given the positive relationship between academic achievement and academic self-concept (see, for instance, Marsh, Chanal and Sarrazin, 2006) explaining the negative compositional effect on achievement as a big fish little pond effect could be justifiable. This hypothesis was, in fact, possible to test and was one of the foci of Study 2 (see section 6.4.1).

Indeed, investigating the mediating role of mathematic self-concept on the negative school compositional effect of average prior achievement on subsequent achievement, I found a positive and a significant indirect effect of school average achievement in year one on students' self-concept in year four via mathematics self-concept in year four (see also section 6.4.1). Nevertheless, the direct effect was also negative and statistically significant (see Table 4.4) so that self-concept only partially mediated this negative compositional effect on subsequent

achievement. With this analysis, I verified the theoretical models on the processes that mediate compositional effects on achievement (e.g. Bandura, 1986). Marsh (1991) also verified an analogous model using data from years ten, twelve and two years after graduation. Again, where I contribute to existing literature (see also Marsh, 1991), is by using structural equation models that correct for measurement error bias (and/or sampling error bias).

A number of other theoretical hypotheses could be made to explain the negative school compositional effect on students' subsequent mathematics achievement that was detected with the analysis of English data in Study1a. However, these suggestions are only tentative given the methodological constraints in my analysis as well as the inherent difficulty in interpreting school compositional effects more generally (see, for instance, section 2.3.4 in the literature review chapter that referes to the distinct and yet inter-related factors underlying the occurrence of school compositional effects and complicate their interpetation).

First of all, the negative effect could have something to do with the schools' strategies and the perspective that teachers take in raising their students' knowledge: in a school where the intake is already high with high achievement in mathematics, teachers will not feel that they need to make an effort to help their students. Rather, they focus on promoting other outcomes for which the student body might be underachieving, for example literacy or science.

Another explanation for this negative compositional effect of school average achievement is that it is confounded with the effect of the pre-school education of the children. For example, work by Tymms, Merrell and Henderson (1997) on the effects of schools on the students' progress in reading and mathematics over the reception year revealed that this progress varied considerably between schools – the variation was much greater than is typically found in school effectiveness studies. Therefore, even though the year one average achievement used for the purposes of Study 1a may be justified as a reasonably good measure of the average school intake, it may be missing potential influences on school average achievement that emerge during the reception year and thus after students start primary school.

Moreover, much of the observed excellence of the students in the early stages of schooling can be attributed to their home environment and to the pedagogic skills of their parents rather than to the school itself. In this way, a student may enter primary school with a good knowledge of mathematics because of the training he or she received at home and not because of his or her actual cognitive skills. Unfortunately, this hypothesis could not be investigated further, since no data on home background characteristics were available.

6.5.5 Methodological Implications for Existing BFLPE research

The methodological implications of this second study of my thesis concern the application of the four models of the 2x2 taxonomy and the impact this has on compositional effects estimates. These have been extensively discussed in Study 1 (see section 6.4 on "Study 1: A Discussion of the Findings" and, specifically, sections 6.4.1 and 6.4.2). Particularly in relation to the negative compositional effects of achievement on subsequent self-concept, namely in relation to investigating the Big-Fish-Little-Pond-Effect-Hypothesis, I note that previous research (e.g. Marsh, 1991) is likely to have underestimated its actual size by not making adjustments for measurement error and sampling error. Thus compositional effects of school average achievement on students' subsequent self-concept are likely to be more negative than they were actually reported to be in previous research.

In my analysis I also demonstrated the usefulness of treating hierarchical structures as a nuisance factor in assessing indirect effects in 2-1-1 designs. This analysis can easily be implemented in Mplus using the "MODEL CONSTRAINT" command. One limitation in applying this approach for the purposes of my thesis was that the estimates of the direct and indirect effects of school compositional effect of average achievement on subsequent outcomes were not directly comparable with the compositional effects estimates obtained using the four models of the 2x2 taxonomy. In this way it was not possible to compare the magnitude of compositional effects obtained with multilevel structural equation models with those obtained

treating the hierarchical structure as a nuisance factor. Although this in itself was not of a focus of the present investigation, it should be taken into account for future studies that may want to make such comparisons. Moreover, using this approach, only adjustments for measurement error are possible and so sampling error bias in estimation is not adjusted for. Although this may not have serious consequences when formative aggregates are involved in the analysis and when the sampling ratio is reasonably high, it can be especially problematic when the focus is on the effects of reflective constructs and when the sampling ratio is small (see section 6.2 on correcting for sampling error in formative and reflective aggregates using doubly latent models).

6.6 Study 3: A Discussion on the Findings

In Study 3, I explored the usefulness of the Regression Discontinuity approach to assess the influence that an extra year of schooling has on students' outcomes and to determine the between-school differences in the size of this effect. In pursuit of this aim I incorporated primary (years four and five) and secondary (years eight and nine) data obtained from TIMSS-95. Regression discontinuity models were used; these were implemented in the multilevel modelling framework. An important focus of Study 3 was to demonstrate how the regression discontinuity approach can be used by educational researchers when seeking to separate the effects of school composition from school effectiveness measures. The focus is on the extent to which measures of school composition (school average achievement) could explain the variability across schools in their added-year effects – the measures of school effectiveness in RD models. Consistent with Study 1 of my thesis, I consider only the potential effect of school composition with primary school data (year four and year five students).

6.6.1 An Extra Year of Schooling does Matter

The impact of one extra year of schooling on students' outcomes using the English TIMSS database has been previously investigated with primary school data (Luyten, 2006). Where I contribute to existing knowledge is that I also consider the absolute effect of schooling with

secondary school data. In my analysis I found positive and significant effects of schooling of students' outcomes suggesting that schools do matter in promoting students' achievement for both phases of schooling. These were approximately the same in magnitude for secondary school students as for primary school students despite the fact that previous research suggests that the effect of schooling decreases with increasing age (see also Cliffordson, 2010). While significant school-to-school variation in their absolute effects could be detected with primary school data, no variation was observed with secondary school data.

6.6.2 Modelling the Relationship between Age and Achievement

By incorporating in my analysis data from two phases of schooling (primary and secondary), it was possible to assess the effect of age on achievement for both primary and secondary school students and compare its magnitude across these two distinct phases. I found positive effects of age on achievement; the effects were larger for primary school data as opposed to secondary school data. This provides evidence that differences in age played a significant role in the achievement scores of students not only in primary school but also in secondary school, albeit smaller for the latter.

6.6.3 No Need for Controls for Background Variables in Regression Discontinuity Models

One of the strengths of the regression discontinuity approach in providing measures of schools' effectiveness as compared to conventional value added modeling is that it does not require adjustments for student background variables and, most importantly, for prior achievement. Provided that the assumption of strict adherence to the cut off age is met, i.e. provided that students begin formal education strictly on the basis of age, then controlling for student-level characteristics in the basic regression discontinuity model should have no impact on the estimates of the absolute effect of schooling and differences across schools in this effect. Consistent with this assumption underlying the RD model, namely the homogeneity of the

characteristics of the population in the lower and in the higher grade, adjustments for the main effects of student background variables in my analysis did not substantially alter the regression discontinuity estimates for the absolute effect of schooling. This was the case with both primary and secondary school data analysis.

6.6.4 Interactions of the Effect of Schooling with Student Background Variables: Investigating Differential School Effectiveness

A significant advantage of using the regression-discontinuity approach to measure the effect of schooling (see also Kyriakides and Luyten, 2009) is the fact that it forces us to make a distinction between the influence that certain variables have on student achievement (through their inclusion in the model as main effects) and their influence on the effect of schooling. The latter been investigated using interaction terms between the year of schooling and the student-level characteristic of interest.

The main effects of student background variables in my data were not of major interest in themselves in my analysis. Therefore I only present them separately, as supplementary materials (see section B.2 in Appendix B): they can be of substantive interest for educational researchers who may want to see how background characteristics affect the achievements of students who are in the same year group.

The absolute effect of schooling in my analysis did not interact significantly with background variables; no evidence for differential school effectiveness was found. Specifically, schools did not appear to be differentially effective for students of different gender, different home environments (number of people at home, number of books in the students' house, parents' ethnicity, availability of home possessions relevant to learning: calculator, computer, study desk for own use, dictionary) or with a different language spoken at home.

6.6.5 Investigating whether the Schools' Composition can Explain Between-School Differences in their Absolute Effects

In explaining between-school variation in their absolute effects with primary school data, school average achievement in mathematics was used. A non-significant effect was identified. Importantly, there was no evidence for positive effects of school average achievement on added-year effects. This suggested that a school with higher average achievement did not necessarily add more "value" to its students during the extra year of school attendance.

Here, it should be noted that, although a positive and significant main effect of school average achievement on the random intercept of the RD model was detected, this should be interpreted very carefully (see "Distinguishing between the effect of school-level variables on random intercepts and random slopes in RD models" and" The effect of school average attainment on added-year effects" under section 5.5.3 in the section describing the results for the RD approach). It suggests only that students who are in the same year (lower or higher) are expected to perform higher. However, this apparent positive effect could be confounded with pre-existing differences in the achievement of the intake of the school in the first place; it does not provide any evidence on the effectiveness on the school itself.

6.6.6 A Potential Advantage of the Use of the RD Approach for School Accountability Purposes

Another interesting finding from my study is the fact that the estimates of the absolute effect of schooling did not substantially alter when school composition variables (school average achievement) were adjusted for to explain between-school variability in the random intercept and the random slope of the RD model. Similar results have been reported, for example by Heck and Moriyama (2010) who also applied the RD approach in the multilevel structural equation modelling to investigate the absolute effect of schooling, differences across schools in their absolute effects and the extent to which variables relevant to the school's composition could

explain these differences (see section 2.5.7 in the literature review). In this way, another advantage of the regression discontinuity approach over conventional value added models of educational effectiveness (see section 2.4.6 in the literature review where I outline these advantages) can be identified: While the schools' value added measures of educational effectiveness, obtained with the conventional multilevel models can substantially alter when school composition variables are included in the analysis (see discussion on Type A and Type B effects in "The use of school compositional effects in value added models of educational effectiveness" under section 2.3.5), added-year effects, obtained using the RD approach for school accountability purposes are more robust to such adjustments. While researchers should be aware of this potential advantage of the RD approach over conventional value added models, they should also always remember the limitations of the RD models (see subsequent section 6.6.8). For instance, with the RD approach the power to detect relative school effects is much lower than traditional value added models, since estimates are based on a restricted number of students – those at the cut off. In this way, the RD approach can be used to complement value added modelling for school accountability purposes – not replace it.

6.6.7 Integrating Multilevel Models with Regression Discontinuity Designs

The originality of Study 3 lies in the fact that it integrates the Marsh, Lüdtke et al. (2009) multilevel structural equation models with the regression discontinuity approach. First of all, adjustments are made for measurement error in individual mathematics achievement –the criterion in regression discontinuity models. Incorporating measurement error in regression discontinuity designs can enhance the accuracy of estimating structural relationships among variables (Raykov and Marcoullides, 2006). In my analysis (see Table 5.7 and Table 5.10) this did not result in significantly different estimates of the absolute effect of schooling; nonetheless, they resulted in slightly smaller standard error estimates for the unstandardized effects.

Next, the possibility of adjusting for measurement and sampling error in the use of school-level aggregates of achievement to explain between-school differences in their absolute effects was explored. Integrating the RD approach with the Marsh, Lüdtke et al. (2009) framework, measurement error was controlled for using multiple indicators while sampling error was corrected for assuming latent aggregation. To be precise, latent interaction models were used in my analysis to investigate the extent to which the absolute effect of schooling was moderated by school-level variables (see section B.1.3 in Appendix B and relevant Mplus syntax in Appendix D.2). The absolute effect of schooling was again found be unrelated to school average achievement; a negative but non-significant moderating effect was retrieved. Hence, based on the findings of the present study, adjustments for measurement and sampling error through latent variable models may not lead to substantially different estimates as compared to those obtained through the use of the multilevel modelling framework, especially when the data are of high measurement and sampling error reliability.

This is the first study ever employed to investigate the impact of aggregate variables at the level of the school to explain between-school differences in their absolute effects. Heck and Moriyama (2010) also used multilevel structural equation models and investigated the impact on students' achievement of a range of school-level variables relevant to the school composition. Nevertheless, they considered only integral variables (i.e. variables at the school-level that could be measured directly) rather than analytical ones - variables which are obtained by aggregating at the higher-level individual-level characteristic (see Lüdtke et al., 2008 for a more detailed description of the difference between these two types of school-level variables). Moreover, in their study the researchers did not consider measurement error in the response variable – students' achievement.

6.6.8 Limitations and directions for further research

6.6.8.1 External validity

The external validity of the estimate of the absolute effect of schooling that is obtained by the regression discontinuity is limited in two respects: Firstly, it refers to only one specific point on the age continuum (Imbens and Lemieux, 2008) so that it should interpret only in relation to that point. Moreover, it relates only to the cohort of students examined and to the specific subject tested. The estimates of the absolute effect of schooling obtained in my analysis are specific only to the education system in England; no generalizations can be made across different systems.

Study 3 provides estimates of the absolute effect of schooling at two different phases of schooling: primary and secondary. This can be considered strength of my study over previous studies which concerned only absolute schooling effects at a specific time point. Where appropriate data are available, regression discontinuity models with multiple cut-off points could be fitted. In this way, it will be possible to investigate the effects of schooling across distinct year groups.

6.6.8.2 The assumption of strict adherence to the cut off

With the regression discontinuity approach, just as with random assignment, it is the exact knowledge of the criteria that determine which students are placed in the lower and which in the upper grades that allows us to obtain unbiased effects of the "impact" of schooling (Shadish, Cook and Cambell, 2002; Rossi, Freeman and Lipsey, 2004). Indeed, exact adherence to the cut-off point is a strict requirement for the approach to yield accurate estimates: Admission to school should be based on only chronological age.

Although some educational systems conform to this principle (e.g. England; data from this country are considered in Study 3), this is not always true. In this way, despite the strengths in applying the RD in educational effectiveness, the nature of schooling in most educational systems (e.g. Australia, United States of America) contradicts the assumption of strict adherence to cut off age: applying the approach to investigate of the absolute effect of schooling will provide only invalid measures of effectiveness.

Cahan and Cohen (1989) explain that there are two ways in which selective misplacement may affect the within-grade regression slopes and, consequently, the between-grade gap representing the schooling effect (see Figure 2.4 representing the absolute effect of schooling): The first relates to the existence of underage and overage children in each grade. The authors explain that the direction of this effect cannot be established a priori since age and selection counteract each other: early or late admission of individuals into schools is not independent of age or intellectual development (see also Cliffordson and Gustafsson, 2011). Brighter students are often accelerated so that they find themselves in the same year group with older groups although, strictly based on their age they should be studying among younger ones. This same set of students often ends up lagging behind, because they have difficulty in catching up with their more mature peers (see also following section on the "Confounders of the absolute effect of schooling") and so more complicated problems with misclassification may result. This can also happen when parents decide to send their children to school at an earlier age because they do not have enough time to take care of them at home. At the same time less bright students are more likely to start school later in time, again possibly resulting in misclassification.

The second way in which selective misplacement may affect the within grade regression is through the existence of "missing" children in each grade. Cahan and Cohen (1989) give a detailed account of how the missing children at both extremes of the age range cause attenuation of the estimated slope leading to underestimated age effect and overestimated schooling effect.

In some cases, there may be a high proportion of misclassified students due to regional variations (e.g. in Australia and the United States) or grade repetition – a system that can often be observed in some educational systems (Luyten, Tymms and Jones, 2009; Luyten and Veldkamp, 2011).

Several ways have been proposed in the literature for dealing with the potentially unreliable assumption of correct classification of the students in grades according to their age. Let me just mention list-wise deletion-when the percentage of non-normal aged children does not exceed 5% (Shadish et al., 2002; Cliffordson and Gustafsson, 2011) - the use of the "fuzzy RD" design and instrumental variable (IV) regression (Cliffordson and Gustaffsson, 2011; Shadish et al., 2002) and addressing the between grades differences in the case of a large proportion of non-normal aged children as a selection-bias problem (Luyten and Veldkamp, 2011). For the purposes of my analysis it was sufficient to list-wise delete the misclassified students from my analysis (see section on "Misclassification: Percentage of delayed/accelerated cases in the samples" in the description of the methododology for Study 3).

6.6.8.3 Confounders of the absolute effect of schooling

Analysis of mathematics data from a number of countries participating in TIMSS 95 (e.g. Austria, Sweden) replicate findings of previous studies that suggest that one year of schooling does exert an effect and that this is about twice as strong as the effect of one chronological year (Cliffordson and Gustaffson, 2011; Luyten, 2006).

A few tentative suggestions can be made as to why the size of the gap for England is so small. One reason might be the fact that the cut-off point could coincide with other significant factors so that a negative bias occurs in the estimate of the absolute effect of schooling. For example, younger students in the higher year may be underachieving while older students in the lower year group may be achieving higher simply due to the prevalence of peer effects: physical differences between the oldest and the youngest children within the same grade, such as maturity or physical size can have an impact on the students' attainment but also on their confidence levels, happiness at school and the risk of being bullied. In a relevant study, Bedard and Duey (2006) showed that the youngest members of the same group of students score lower than the older ones for consecutive years and suggest that they are even less likely to attend university. In this way, students who start school at an earlier age can be at a cognitive disadvantage relative to their older peers: they are more susceptible to emotional difficulties, learning disabilities and school retention and are less advanced in many cognitive and behavioural skills that facilitate reading acquisition and general learning.

To understand this better, I refer the reader to Figure 2.4 in the literature review where the RD estimates are displayed graphically. Older students in the lower year group (corresponding to the right end of the regression line that describes the relationship between age and achievement for the lower year group) perform higher because of their physical and intellectual advantage as compared with their peers. At the same time younger students in the higher year group (corresponding to the left end of the regression line that describes the relationship between age and achievement in the higher year group) perform lower. In this way the gap - indicative of the absolute effect of schooling is estimated smaller.

Another explanation of this phenomenon – absolute effects in England being estimated smaller as compared with other countries -- can be the results of the fact that in other countries the cut off rule is not followed as strictly so that a higher percentage of brighter students find themselves in the upper grade and a smaller percentage of less able students find themselves in

the lower grade (see also previous section on "The assumption of strict adherence to the cut off"). In this way the gap – that represents the difference between the mean achievements of the students at the cut-off who are in the lower grade from that of the students who are in the higher grade – can be overestimated. Even if these children are excluded in a list-wise manner from the analysis – just as I did in my analysis ("Misclassification: Percentage of delayed/accelerated cases in the samples" under section 5.3.1 in methods) – the result will again be a positive bias in the gap. One could thus claim that the bias in the size of the gap caused by misclassification is positively correlated with the percentage of misclassification in the data: a higher proportion of misclassification results in more positive bias in the estimated effects of schooling (larger absolute schooling effects).

6.7 The Use of Early Primary School Data in the Assessment of the Magnitude of the Potential Effect of School Composition

The first two studies of my thesis are concerned with the effects school-level aggregates of prior achievement in mathematics during the early primary school years -first four years of primary education. Even though Study 3 investigate the absolute effect of schooling both with primary and secondary school data, it only considers the impact of school composition on the schools' added-year effects using solely primary school data (from year four and year five of primary education). Studying this age range is fairly uncommon in the existing literature; studies tend to focus on older children and on outcomes at the end of the primary stage (e.g. Lauder, Kounali, Robinson and Goldstein, 2010; Kyriakides and Tsangaridou, 2008). Nonetheless, it is crucial to obtain an insight into the early stages of schooling. This is particularly the case if we consider the compelling evidence indicating that early schooling has a considerable impact on the subsequent academic development of a child (Darlington, Royce, Snipper, Murrey and Lazar, 1980; Kyriakides, 2008; Lazar, Darlington, Murray, Royce, Snipper and Ramey, 1982; Hess, Holloway, Dickson and Price, 1984; Stevenson and Newman, 1986, Goldstein and Sammons, 1997).

6.8 Methodological Restrictions in the Assessment of School Compositional Effects

In the literatue (Gray, Jesson and Sime, 1990; Harker and Tymms, 2004), two facets of under-specification have been identified in studies investigating compositional effects (see also section 2.3.8 in the literature review entitled "The two facets of under-specification at level 1: the omission and the mis-measurement of relevant variables"): (i) the existence of measurement error in the individual-level predictors and (ii) the failure to control for an adequate number of level 1 covariates. Both of these factors have been shown to lead to phantom compositional effects with conventional compositional analysis models.

In my analyses concerning compositional effects on students' mathematics outcomes with Study 1, I demonstrate and propose ways around only the first issue—that related to a predictor's reliability. In the first place, the accurate measurement of the predictors included in a contextual analysis model is essential. Even if all the relevant covariates are included in the model, the derived estimates will not be valid unless adjustments are made for measurement error for each and every one of them. However, due to inadequate controls of student-level background characteristics my analysis is restricted and does not allow me to make extensive theoretical inferences. Nevertheless, the negative effect of school average ability detected in this first analysis – which is concerned with the English data - could potentially be achieved even in a more elaborate statistical analysis. Phantom effects, the result either of failure to include appropriate covariates, or of measurement error in the individual-level scores, are a positive bias to estimates of compositional effects. Thus, correcting for phantom effects due to under-representation typically makes apparently positive compositional effects less positive (or non-significant or even negative, depending on the size of the correction). The estimated compositional effects in the first study are negative; hence, the addition of further covariates is unlikely to make the compositional effects more positive.

Juxtaposing compositional analysis models (Study 1 and Study 2) and the use of the RD approach in investigating the effect of school composition, clearly the latter provides more robust estimates with respect to the two facets of under-representation at level 1 (see section 1.7 in the introduction and section 2.3.8 in the literature review). Once the main assumptions underlying the RD approach are fulfilled, there is no requirement for controls for student background variables other than age. Moreover, measurement error in student-level achievement does not bias the RD estimates - individual achievement is the criterion in RD models. Hence, the two faces of under-representation at level 1, which relate to traditional compositional analyses models are no more relevant to RD designs.

6.9 Theoretical Implications

A central focus of educational researchers and policy makers around the world is the impact of the polarisation of school intakes on students' individual achievement. This is the extent to which the clustering of students together in high or low ability settings, either on purpose, by selection practices, or de facto, due to social or residential segregation and travel limitations, has a positive or a negative effect on the students' individual progress. This issue is quite relevant to England which experiences relatively high levels of segregation.

The results of the first two studies of the present investigation suggest that, on average, there are no benefits, and that there are even negative consequences, in terms of mathematics self-concept (Study 2) or, even mathematics achievement (Study 1) for students who attend a school with a high-achievement intake. With Study 3, a negative and non-significant relationship was found between school average achievement and added-year effects. My findings call into question the supposed advantages of attending higher achievement schools. Although the potential disadvantages of attending higher achievement schools, as indicated by my analyses, may not generalize to all high achievement schools and to all students, at least for some children, the impression they will have of themselves as poor students in a competitive

environment may be more detrimental than the potential benefits of attending an academic selective institution. Indeed, existing research (e.g. Hattie, 2002; see also section 2.3.3 in the literature review chapter) suggests that the apparent positive effects of practices such as tracking and ability grouping which are often inferred based on the estimation of compositional effects of achievement students' subsequent academic and affective accomplishments can often prove to be deceptive.

Given the small and even negative compositional effects of achievement that have been detected in my analyses, it is clearly important for the educational community to identify other effectiveness factors. These could be applied by schools or teachers (see, e.g., Creemers and Kyriakides, 2010) in order to enhance the academic progress of their students independently of their initial educational outcomes. Indeed, recent theories exploring the different facets of educational effectiveness (e.g. Creemers and Kyriakides, 2008; 2010) suggest that school processes and practices can be more important than school composition in explaining students' outcomes. Thus, policy makers and people responsible for educational reforms should place more emphasis on factors relevant to the implemented policies and the practices that the schools adopt for instruction and evaluation (e.g. school policy on teaching; questioning skills of teachers; opportunity to learn; the creation of a supportive environment).

6.10 Directions for Future Research

The methodology demonstrated can also be used to evaluate practices that are widely implemented around the world and that may have implications on the extent to which schools become segregated in terms of their pupils' ethnicity and income. For instance, a central focus of the current educational policies in England has been to expand the parent choice over where their children go to school (Gibbons, Machin and Silve, 2011). This policy is expected to generate competition between schools and, in the schools' effort to attract students, standards will be raised. (Dearden and Vignoles, 2011). Gibbons et al. (2011) claim that despite the

potential performance advantages that such practices may offer, they may lead to increasing social polarization: Schools may become more segregated in terms of the pupils' ethnicity and income while academically able and less able children may become segregated into different schools. It is important that policy makers should exercise caution in implementing such policies in order not to impact negatively on the very outcomes that they intend to maximize. According to the findings of the present study (negative effect of school average ability on achievement), relevant practices can have detrimental effects on students' achievement. However, based on the findings my thesis alone, I cannot make valid judgments on such educational policies (see section 6.8 on methodological restrictions), but I suggest that educational research use the Marsh, Lüdtke et al. 2x2 taxonomy of models in order to address such questions with appropriate datasets. Subsequent research could even use the regression discontinuity approach to explore these issues, investigating, for example, the simultaneous effects of the schools' composition and the schools' policies and instructional practices (see for example, Heck and Moriyama, 2010).

6.11 Conclusion

Many of the older and ongoing criticisms that pervade the field of educational effectiveness (Coe and Fitz-Gibbon, 1998; Tizard et al., 1980) relate to how failure to take into account measurement error in the underlying educational data may result in invalid inferences. Measurement error in students' achievement scores has proved to be particularly problematic in compositional analysis seeking to verify the impact of school (or class) - level aggregates (e.g. school average achievement, school average socio-economic status) on students' outcomes. Specifically, measurement error in the individual-level measure on which aggregation is based has been shown to lead to phantom compositional effects due to positive bias in the effect of the higher-level aggregate.

In educational effectiveness, school-level aggregates are used, for instance, in value added models of educational effectiveness to explain variability in relative school effects. My thesis proposed the use of the Marsh, Lüdtke et al. (2009) framework, and in particular the use of the manifest latent and doubly latent approach from the Marsh, Lüdtke et al. (2009) 2x2 taxonomy of models, for the assessment of the magnitude of school compositional effects; the models can control for measurement error bias (and/or sampling error bias) in compositional effects estimates. Using mathematics achievement primary school data from England and from Cyprus (Study 1), I demonstrated how these models correct for the positive bias in the estimates of compositional effects. Interestingly enough, non-significant and even negative school compositional effects of prior mathematics achievement on students' subsequent achievement in mathematics were retrieved – in contrast with conventional wisdom.

The Marsh, Lüdtke et al. (2009) partial and doubly latent models were originally developed in the field of educational psychology, driven by research on frame-of-reference effects and especially by the BFLPE hypothesis. The BFLPE was also a major substantive focus of my thesis – I verified this phenomenon for students in year one and year four of primary education (Study 2). Bridging Study 1 and Study 2, I demonstrated that the negative school compositional effect of prior mathematics achievement on students' subsequent mathematics achievement, as detected in Study 1 was, in part, mediated by academic self-concept. Hence, this negative school compositional effect could be attributed to the prevalence of BFLPEs with these data.

Lastly, embracing a perspective of school accountability which is entirely different to that of conventional value added modelling, I investigated the effectiveness of schools in terms of the "value" that they add to their students during one extra year of schooling (Study 3). To this end, I applied the Regression Discontinuity (RD) approach to English primary and secondary TIMSS-95 mathematics achievement data. I demonstrated that the estimates of school effectiveness (absolute effect of schooling and added-year effects) obtained in my

analysis do not alter substantially when student background variables controlled in the RD model or when unreliability due to measurement error in student-level achievement data is adjusted for. Moreover, I explored the potential to use the RD approach in assessing the effects of school composition on students' outcomes. With the primary school data (years four and five) in particular, I considered the extent to which added-year effects (the measures of school effectiveness with RD models) correlated with school average achievement: a non-significant effect was retrieved both when conventional multilevel models were used and when multilevel structural equation models were incorporated; the latter adjusting for measurement error in student-level achievement and for measurement and sampling error in the school-level aggregate.

My findings relate to on-going debates underlying the field of educational effectiveness on the extent to which school composition can influence students' outcomes. The methodology proposed can be used in future studies to investigate relevant substantive issues (e.g. the impact of segregation and polarization of intakes on students' outcomes).

Appendix A: School Compositional Effects and the Marsh, Lüdtke et al. (2009) Framework

The conventional statistical approach for the investigation of compositional effects is multilevel modelling. This methodological approach involves the use of single manifest indicators at the level of individual student. Moreover, when investigating compositional effects in this framework, the aggregate score is computed as the observed average of individual-level observations within each group. In this technical Appendix I present the basic multilevel model for investigating compositional effect. I explain further the notion of measurement error and sampling error. Using mathematical formulas I outline how educational researchers, when using multilevel modelling, fail to control (i) for measurement error in individual and group-level variables and (ii) for sampling error related to the aggregation of level 1 constructs to form the group-level constructs. The first part of this Appendix (section A.1.1) resembles section 2.2.1 in the literature review chapter ("Compositional Analysis Models"). This is to enable the reader to make the linkage between the mathematical derivations presented here and the more theoretical aspects presented in the literature review.

A.1 Multilevel Statistical Models for the Investigation of Compositional Effects

A.1.1 The Basic Compositional Effects Model

Suppose that there is a two-level structure in the data, with individuals nested within-groups (e.g. students nested into schools) and that an individual-level variable X (e.g. prior attainment; as in my analyses) predicts a dependent variable Y (e.g. subsequent attainment in the context of educational effectiveness and value added modelling and self-concept in the context of self-concept research). Then the two-level model, conventionally used for the investigation of compositional effects, can be expressed in the following way (the notation is adopted by Snijders and Bosker, 2004):

Level 1: $Y_{ij} = \gamma_{oj} + \gamma_{10}(X_{ij} - \overline{X}_{.j}) + R_{ij}$ (A.1.1)

Level 2: $\gamma_{0j} = \gamma_{00} + \gamma_{01}\overline{X}_{.j} + U_{0j}$ (A.1.2)

The equations given in relationship (A.1.2) can be combined with equation (A.1.1) giving:

$Y_{ij} = \gamma_{00} + \gamma_{10}(X_{ij} - \overline{X}_{.j}) + \gamma_{01}\overline{X}_{.j} + U_{0j} + R_{ij}$ (A.1.3)

In relationship (A.1.1), Y_{ij} is the outcome for person i in group j. The individual-level predictor X_{ij} is centred on the group mean $\overline{X_{.j}}$ (group-mean centring[5]). In relationship (A.1.2) the intercept γ_{0j} is the dependent variable. Relationship (A.1.3) describes a random intercept model—only the intercept is allowed to vary across groups.

Group mean centring the predictor variable implies that γ_{01}, the slope relating the random variable $\overline{X_{.j}}$ to the intercepts from level 1 equation is the between-group regression coefficient, describing the relationship between the aggregates $\overline{X}_{.j}$ and $\overline{Y}_{.j}$. The fixed effect of the individual-level variable X on Y, (γ_{10}) is the within-group coefficient describing the relationship between X_{ij} and Y_{ij} within each group (For a more detailed discussion on within- and between- regressions see Snijders and Bosker, 2004). To obtain the compositional effect—that is, the effect of the group average $\overline{X}_{.j}$ on Y_{ij}, the difference γ_{01} - γ_{10} should be used: a compositional effect is present if γ_{10} is significantly different from γ_{01}.

---

[5] An equivalent way to specify a compositional analysis model would be to use grand mean centring of the predictor by subtracting from each level 1 observation the grand mean of the level 1 predictor. Lüdtke et al. (2008, pp. 206-207) explain how the compositional effect can be obtained by the grand mean centring approach, and the equivalence between the grand mean centring approach and the group mean centring approach.

The parameter γ_{00} represents the overall mean of the outcome across all individuals and across all groups, while the quantity U_{0j} represents the residual at the group-level—this is denoted by allowing only the indicator for the group, j, to appear in the subscript of the residual, showing that the residual can only vary between groups. Its variance represents variability between groups that is not explained by the explanatory variable. The residuals at the within level are given in the relationship (A.1.1) by R_{ij}. In an analogous way, they indicate the extent of variability between individuals not explained by the explanatory variable. The residuals U_{0j} and R_{ij} are assumed to be independent from each other and are normally distributed, with mean zero and variance τ^2 and σ^2 respectively.

A.1.2 Defining the Intra Class Correlation Coefficient

The Intraclass Correlation Coefficient ($ICC, ICC1$) is often used in compositional effects models to quantify the between-group variability of a specific variable. It represents the proportion of the variation in a given variable that is accounted for by the differences between clusters. The mathematical formula for the ICC can be given as:

$$ICC = \frac{\tau^2}{\tau^2 + \sigma^2} \quad \text{(A.1.4)}$$

where τ^2 is the between-group component of variance and σ^2 is the within-group component of variance of the variable of interest.

A.1.3 The ICC and its Role in Estimating School Effects in Value Added Models

Within the value added modelling framework in school effectiveness research, the groups are defined by the schools. Value added models should at least control for prior achievement, this is variable X_{ij} in relationships A.1.1- A.1.3.Then, the deviation of each school from the overall mean, given X , which is the level 2 residual, U_{0j} , is taken to be an estimate of the school's value added.

The ICC , as defined in relationship A.1.4, is closely related to the definition of school effects (the proportion of variance in the outcome variable not explained by background variables attributed to the group) and the way they are conceptualized in multilevel modelling framework. School effects in multilevel modelling are given by dividing the residual variance at the school-level after adjustments for intake by the total variance in the students' achievement scores. This is a slightly different quantity to the ICC , in which the denominator in the ratio is actually the residual variance at the student-level after adjustments for intake, although the actual ICC can also be used for quantifying school effects.

A.2 Consequences of Unaccounted Measurement and Sampling Error on the Estimation of Compositional Effects

Traditionally, multilevel models (MLM) use single manifest indicators at level 1, appearing in the model as X_{ij} and Y_{ij} , and compositional effects models use level 2 variables, $\overline{X}_{.j}$, computed as the average of the level 1 variables within each group. In this way, there is no control for any measurement error in the level 1 or level 2 variables or for sampling error related to the aggregation of level 1 constructs to form level 2 constructs. In the section that follows I demonstrate the bias that can become evident in the estimates of compositional effects due to

measurement and sampling error. These are quantified through the use of mathematical formulas, as these were derived by Lüdtke et al. (2011) and Fletcher (2012).

A.2.1 The Notion of Measurement Error Reliability in Classical Test Theory

The fundamental idea in Classical Test Theory is that an observed score is decomposed into a true score and an error component (Lord and Novick, 1986; Lord, Novick and Birnbaum, 1968) in the following way:

$X = T + E$ (A.2.1),

where X denotes the observed score of the individual, T denotes the true score and E is the random component.

The accuracy of a measurement in CTT can be quantified by the notion of reliability. Reliability is closely related to the ability to replicate a measurement (Brennan, 2010). The extent to which a measure is reliable depends on the degree of consistency in the individual scores over replications of a measurement procedure.

Scale reliability denotes the proportion of observed score variability that is attributable to true score variability across individuals; reliability is higher when the occurrence of error is lower (Creemers, Kyriakides and Sammons, 2010). There exist several different methods for computing reliability and each conceptualizes and operationalizes reliability in a different way; nevertheless they all share this fundamental definition. Based on equation (A.2.1), the reliability coefficient of a measurement device as the ratio of the true score variance (σ_{T}^2) divided by the observed score variance (σ_{χ}^2), that is equal to the sum of the true score variance and the error variance ($\sigma_{\mathrm{T}}^2 + \sigma_{\mathrm{E}}^2$).

A.2.2 Modelling Multilevel Measurement Error and Sampling Error

Lüdtke, Marsh, Robtizsch and Trautwein, (2011) extended the classical test theory to the multilevel case, distinguishing between random error that varies systematically across groups and measurement error at the individual-level. They proposed the decomposition of observations of level 1 units (nested in level 2 groups) in the following way:

$$X_{ij} = \mu_x + U_{xj} + U_{xij} + R_{xj} + R_{xij} \quad \text{(A.2.2)}$$

In the above formula, μ_x corresponds to the grand mean of all observations across all people and all clusters, U_{xj} is the true score at level 2 with mean zero and variance τ_{χ}^2 and the parameter U_{xij} is the true score at level 1with mean zero and variance σ_{χ}^2, R_{xj} is the error score at level 2 and R_{xij} is the error score at level 1 with variance $\tau_{\chi,e}^2$ and $\sigma_{\chi,e}^2$ respectively. The latent variables in (A.2.2) are all assumed to be independent from each other.

When individual-level variables are aggregated up at the school-level, then the school-level variable is given by:

$$\overline{X}_{.j} = \mu_x + U_{xj} + \overline{U}_{x.j} + R_{xj} + \overline{R}_{x.j} \text{(A.2.3)}$$

In this formula the terms μ_x, U_{xj} and R_{xij} can be interpreted in the same way as formula (A.2.2). The term $\overline{U}_{x.j}$ denotes the average value of the true score at level 1 (U_{xij}) while $\overline{R}_{x.j}$ denotes the average value of the error scores at level 1(R_{xij}).

The random measurement error (see Lüdtke et al., 2012): at the level of the group is comprised of the last two independent components in relationship A.2.3. The first component, R_{xj}, results purely from influences at level 2 that may distort the measurement of the group-level construct (the group specific influences that I have discussed in section 1.3 of my introduction in relation

to level 2 measurement error). The second component, $\overline{R}_{x.j}$, is essentially measurement error at the level of the student aggregated up at the level of the school. The larger the number of individuals within each school, the smaller the variance of this error component.

Even when measurement error is taken into account, the true value of the level 2 construct, when obtained by aggregating individual-level observations, is still not purely represented by the remaining terms in equation (A.2.3). There exists another source of error, in equation (A.2.3); this is denoted by $\overline{U}_{x.j}$ and results from the finite number of individuals that are used to assess the value of the group-level construct. The variance of this error component tends to zero as the number of individuals with each cluster becomes larger and larger tending to infinity. This second source of error occurring in group-level constructs is sampling error. Thus, according to (A.2.3) sampling error is conceptualised as the error that is due to observing only a finite sample of an infinite population.

Now, given the notation introduced in this paragraph, the true relationship between the outcome variable Y and the true values of the explanatory variables is given by:

$$Y_{ij} = \mu_Y + \beta_{between} U_{xj} + \beta_{within} U_{xij} + \delta_j + \varepsilon_{ij} \quad \text{(A.2.4)},$$

where μ_Y is a grand mean for the outcome, $\beta_{between}$ is the true between group effect, β_{within} is the true within group effect, δ_j is the residual at the level of the school and ε_{ij} is the residual at the level of the individual student.

Note now that the variables $\overline{X}_j$ and $X_{ij} - \overline{X}_{.j}$ that are being used in the conventional multilevel modelling approach (see previous section and equations A.1.2-A.1.3) are the observed measures of the true scores U_{xj} and U_{xij}. Thus, no account is taken of measurement error in the explanatory variable or for sampling error in the cluster mean: the coefficients γ_{01}

and γ_{10} are estimates of the regression coefficients $\beta_{between}$ and β_{within} respectively and γ_{01} - γ_{10} is an estimate of the true value of the compositional effect $\beta_{between} - \beta_{within}$. The prevalence of measurement and sampling error in the observed scores induces bias in these estimates; I expand on this in the next section.

A.2.3 Reliability and Bias due to Sampling Error

Based on the above equation the reliability of the cluster mean $\overline{X_{.j}}$ as given by (Searle, Casella and McCulloch, 1992) is:

$$\text{Rel}(\overline{X_{.j}}) = \frac{\tau_x^2}{\tau_x^2 + \tau_{x,e}^2 + \frac{\sigma_\chi^2}{n_j} + \frac{\sigma_{x,e}^2}{n_j}} \quad \text{(A.2.5)}$$

In relationship (A.2.5) n_j is the number of people in each group; the average cluster size can be used if not all groups are of the same size.

If we assume that there exists no measurement error at level 1 and level 2 with $\tau_{x,e}^2 = \sigma_{x,e}^2 = 0$, then:

$$\text{Rel}(\overline{X_{.j}}) = \frac{\tau_x^2}{\tau_x^2 + \frac{\sigma_\chi^2}{n_j}} = \frac{n_j.ICC}{1 + (n_j - 1)ICC} \quad \text{(A.2.6)}$$

The above relationship is defined as the ICC 2 as opposed to ICC 1or ICC in the literature (see Bliese, 2000) and it is used to assess sampling error reliability for reflective constructs.

Lüdtke et al. (2008) derived a formula for the bias that arises in the estimate of the compositional effect by the doubly manifest approach:

$$E(\gamma_{01} - \gamma_{10}) - (\beta_{between} - \beta_{within}) = (\beta_{within} - \beta_{between}).\frac{1}{n}.\frac{(1 - ICC)}{ICC + (1 - ICC)/n} \quad \text{(A.2.7)}$$

The above relationship suggests that, when adjustments are not made for sampling error, the compositional effect is underestimated if $\beta_{within} < \beta_{between}$ (suggesting a positive compositional effect) but it is overestimated when $\beta_{within} > \beta_{between}$ (implying a negative compositional effect). Moreover, when only a small number of individuals are sampled from within each group and when ICC is smaller, suggesting greater variability in the scores of the individuals within each group, the bias in the compositional effect can be substantial.

A.2.4 Reliability and Bias due to Measurement Error

In the previous discussion the focus was on sampling error in the absence of measurement error. In the discussion that follows I provide some measures for the reliability of observed individual-level and group-level scores due to measurement error. The assumption is that there is no sampling error in the measurement. The total score is assumed to be obtained averaging across the scores of individuals across different items or variables. Thus, the assumption is that the construct is measured by multiple indicators. Analogously with (A.2.2) each indicator can be decomposed into a within and a between part. We can further assume that the indicators are mean-centred so that the grand mean is zero and that the factor loadings are all one:

$$X_{kij} = U_{xij} + R_{xkij} + U_{xj} + R_{xkj} \text{ (A.2.8)}$$

In the above relationship all the components can be interpreted just as the components of equation (A.2.2) but there exists an additional index: k is used as an index for indicators, i for individuals and j for groups. The assumption is that the true score is measured by $k = 1,..., K$ indicators; it is estimated as the mean across the items:

$$\overline{X_{.ij}} = \frac{\sum_{k=1}^{K} X_{kij}}{K} \quad \text{(A.2.9)}$$

The finite numbers of items that are used to measure this level 1 score are assumed to be a subset of potentially infinite items that could have been used to provide a perfectly reliable measurement. The formula for the reliability due to sampling of items is analogous to the formula for the reliability due to sampling of people. Following Lüdtke, Marsh, Robitzsch and Trautwein (2011) it can be defined both at level 1 and level 2 by:

$$\mathrm{Re}\, l_{L1}(\overline{X_{.i}}) = \frac{\sigma_x^2}{\sigma_\chi^2 + \sigma_{\chi,e}^2 / \mathrm{K}} \text{ (A.2.10)}, \quad \mathrm{Re}\, l_{L2}(\overline{X_{.ij}}) = \frac{\tau_x^2}{\tau_x^2 + \tau_{x,e}^2 / K} \text{(A.2.11)}$$

Note the similarity between the above equations and equation (A.2.6).

Based on the reliability estimates from equations that display the reliability of $\overline{X_{.j}}$ due to sampling error, the reliability of X_{ij} due to measurement error (A.2.10) and the reliability of $\overline{X_{.j}}$ due to measurement error (A.2.11), Lüdtke et al. (2011) show that the bias in the estimation of the within effect β_{within} and the between effect $\beta_{between}$ may be assessed by the following relationships:

$$E(\gamma_{10} - \beta_{within}) = -\beta_{within}(1 - \mathrm{Re}\, l_{L1}) \text{ (A.2.12)}$$

$$E(\gamma_{01} - \beta_{between}) = -\beta_{between}(1 - \mathrm{Re}\, l_{L_2} \cdot \frac{1}{1 + r.b/n}) + \beta_{within.} \frac{b}{n} \mathrm{Re}\, l_{L_2} \frac{1}{1 + rb/n} \text{ (A.2.13)}$$

In equation A.2.11, $r = \mathrm{Re}\, l_{L2} \big/ \mathrm{Re}\, l_{L1}$ is defined as the ratio of the reliabilities at level 1 and level 2 and $b = \sigma_{\zeta}^2 \big/ \tau_x^2$ is the ratio of the variance within groups to the variance between groups while n is the number of L1 units within each L2 unit.

The bias for β_{within} depends only on the measurement error reliability, higher unreliability at level 1 attenuating the estimated effect. The bias for $\beta_{between}$ depends on:

(i) The reliability of measurement of X with higher reliability leading to smaller bias
(ii) The group size n with larger group size leading to smaller bias and
(iii) The ICC of the predictor with the bias decreasing as the ICC increases.

A.2.5 Considering all Possible Types of Misspecification

The discussion until now has been focused on aggregations based on infinite sampling. The type of aggregation (formative or reflective) does not affect the expected value for the bias in the estimation of β_{within} as this is given in equation (A.2.12). Nevertheless, depending on whether an infinite or finite sampling process should be followed, misspecification due to sampling error can occur in a number of different ways. Each different type of misspecification implies a distinct value for the expected bias in the estimation of $\beta_{between}$. Fletcher (2012) derived mathematical formulae for the estimates of $\beta_{between}$ corresponding to all the possible combinations between the correct sampling process and the assumed sampling process.

In Table A.1 the term “infinite” refers to a reflective aggregation process, the term “cluster” is used when the individuals within each group (e.g. class or school) form the population of interest and the term ”sub-sample” when only a part of the total number of individuals within the cluster is sampled. Whenever “cluster” sampling is assumed, the size of the cluster is denoted by n and whenever a “sub-sample” is selected from each cluster the size of this smaller sample is given by r. The measurement error reliability for infinite sampling process is denoted

by the notation $\mathrm{Re}\, l_b^{\infty}$. This is given by equation (A.2.11) assuming a single indicator ($k = 1$). Fletcher generalized the expression (A.2.9) that assumes that clusters ought to be modelled as samples from hypothetical infinite populations, so that it relates to the case where the data are modelled as samples from clusters of size n :

$$\mathrm{Re}\, l_b^n = \frac{\tau_x^2 + \sigma_x^2 / n}{\tau_x^2 + \tau_{x,e}^2 + (\sigma_x^2 + \sigma_{x,e}^2)/n} \quad \text{(A.2.14)}$$

In Table A.1, r is replaced with n when a sub-sample of size r is selected from the total population.

In concluding, in the first section of this Technical Appendix (section A.1) I gave the basic compositional effects model being used in the traditional multilevel modelling framework and I explained how the ICC can be used to estimate school effects in value added models of educational effectiveness. The focus of my thesis is mainly on compositional effects estimates; formulas for the bias in the estimated effects due to unreliability in the level 1 and level 2 variables involved in the analysis are given in section A2. Two different sources of error have been considered: Measurement error and sampling error.

Table A. 1. Expression for the expected value of the between group effect when a model is mis-specified with regard to measurement error on the explanatory variable. The table cross-classifies whether sampling error occurs and whether it is taken into account in the modelling procedure.

Assumed sampling process	**Underlying population**	
	Infinite	**Cluster**
Infinite	$\mathrm{Re}\, l_b^{\infty} \beta_{between}$	$\mathrm{Re}\, l_b^{\infty}[\beta_{between} + (\beta_{between} - \beta_{within}) \frac{\sigma_x^2 / n}{\tau_\chi^2}]$
Cluster	$\mathrm{Re}\, l_b^{n}[\beta_{bewteen} + (\beta_{within} - \beta_{between}) \frac{\sigma_x^2 / n}{\tau_\chi^2 + \sigma_\chi^2 / n}]$	$\mathrm{Re}\, l_b^{n} \beta_{between}$
sub_sample	$\mathrm{Re}\, l_b^{r}[\beta_{between} + (\beta_{within} - \beta_{between}) \frac{\sigma_x^2 / r}{\tau_\chi^2 + \sigma_\chi^2 / r}]$	$\mathrm{Re}\, l_b^{r}[\beta_{between} + (\beta_{within} - \beta_{between}) \frac{(1/r - 1/n)\sigma_x^2}{\tau_\chi^2 + \sigma_\chi^2 / r}]$

Source: Fletcher, J. (2012) Regression Artefacts and the Measurement of Value Added in Schools. Confirmation of Status. Oxford: Department of Education.

Note. In this table β_{within} represents the within group effect, $\beta_{between}$ represents the between effect, $\mathrm{Re}\, l_b^{\infty}$ is used to denote the measurement error reliability for an infinite sampling process while $\mathrm{Re}\, l_b^{k}$ is used to denote measurement error reliability when cluster sampling is assumed and when a sample of size k is sampled from a cluster of size n. When the total number of observations is used, k is replaced with n. The parameter σ_χ^2 is used to denote the within group variance of the true value of the explanatory variable X while the parameter τ_χ^2 represents the between group variance of the true value of the explanatory variables.

A.3 A Technical Presentation of the Models from the 2x2 Taxonomy

Marsh, Lütke et al. (2009) set out four contextual models that can accommodate for measurement error and for sampling error. These models are incorporated into my study to investigate compositional effects. Here I expand this description, by providing the mathematical formulas for each as these are adopted by Lüdtke et al. (2011). The reason I decided to include this mathematical form of the models was because I believed it could help the reader understand how the different variables involved in the analysis are calculated and how the four models differ from each other. For all four models, the level 1 (L1, level 1, individual-level) units are the students, indexed by i, who are viewed as being nested within the level 2 (L2, level 2, group-level) units, the schools, indexed by j.

A.3.1 Model 1: The Doubly Manifest Approach

The dependent variable Y, which, in the present analysis is either the students' mathematics achievement or the students' mathematics self-concept, is obtained by averaging across the L indicators for each student in school j: $\overline{Y}_{.ij} = \frac{1}{L}\sum_{l=1}^{L} Y_{lij}$. Adjustments are made for the predictor X given by the average of the score of each student across the K indicators corresponding to the corresponding student-level measure (in my analysis student mathematics achievement) $\overline{X}_{.ij} = \frac{1}{K}\sum_{k=1}^{K} X_{kij}$. The school-level (level 2) average attainment is calculated by averaging across the prior attainment of the students in the school j $\overline{X}_{..j} = \frac{1}{K.n_j}\sum_{l=1}^{n_j}\sum_{k=1}^{K} X_{kij}$, where n_j is the number of students in school j.

Model 1, given by the structural equation:

$$\overline{Y}_{.ij} = \beta_0 + \gamma_{10}(\overline{\mathrm{X}}_{.ij} - \overline{\mathrm{X}}_{..j}) + \gamma_{01}\overline{\mathrm{X}}_{..j} + \delta_{0j} + \varepsilon_{ij} \quad \text{(A.3.1)}$$

The parameter β_0 is the grand mean intercept. Note that in the equation above, group mean centring is used for the predictor variable X. Therefore, γ_{10} is the coefficient that describes the relationship between the prior and the final attainment within schools (within group regression coefficient) and γ_{01} is the between school regression coefficient indicating the relationship between the school average prior attainment and the mean final attainment. The compositional effect β_{cont}, which is the effect of the average prior attainment on the final attainment is given by the difference between the between group effect γ_{01} and the within group effect β_{within} The school-level residual δ_{0j} captures differences in the final attainment, after controlling for the prior attainment and the average prior attainment, between schools. In value added terminology, this residual gives the value added score of the school. Then the individual-level residual ε_{ij} captures deviations of the individual scores from this school mean. These residuals are assumed to be normally distributed with an expected value of zero and are uncorrelated with each other.

Model 1 ignores measurement error due to sampling of items and sampling error due to the sampling of the students from within schools. It is therefore labelled the "doubly manifest" approach**.**

A.3.2 Model 2: The Manifest Latent Approach

The second model of the 2x2 taxonomy, Model 2 controls for unreliability in the average prior attainment due to sampling error (latent approach) but not for unreliability due to measurement error (manifest approach). Lüdtke et al. (2011) refer to this approach as the Multilevel Latent Covariate (MLC) approach as opposed to the Multilevel Manifest Covariate (MMC) approach expressed by Model 1.

In the MLC approach the observed scores of the predictor variable X at the individual-level can be decomposed into "unobserved components that can be viewed as latent variables" (Lüdtke et al., 2011, p.450) and that represent the true values of the variable at the within and the between level:

$$\begin{aligned} Y_{ij} &= \mu_y + U_{yj} + R_{yij} \\ X_{ij} &= \mu_x + U_{xj} + R_{xij} \end{aligned} \quad \text{(A.3.2)}$$

In the above relationships μ_x is the overall mean of X , U_{xj} are the group specific deviations and R_{xij} are individual deviations. The decomposition of Y is interpreted in a similar way. The interest is in the relationship between the latent variables at the individual and at the school-level:

$$\begin{aligned} R_{yij} &= \mu_y + \gamma_{10} R_{xij} + \varepsilon_{ij} \\ \mathrm{U}_{\mathrm{xij}} &= \mu_x + \gamma_{01} U_{xij} + \delta_{0j} \end{aligned} \quad \text{(A.3.3)}$$

Thus, if we combine equations (A.3.2) and (A.3.3), the structural equation for Model 2 is given as follows:

$$Y_{ij} = \mu_y + U_{yj} + R_{yij} = \mu_y + \beta_2 U_{xij} + \delta_{0j} + \gamma_{10} R_{xij} + \varepsilon_{ij} \quad \text{(A.3.4)}$$

If we compare equation (A.3.1) with equation (A.3.4), it can be seen that the latent individual-level deviation U_{xij} in equation (A.3.4) is approximated by the observed individual-level deviations $\overline{X}_{.ij} - \overline{X}_{..j}$ in equation (A.3.1) while the latent group-level deviation U_{xj} is approximated by the observed aggregate $\overline{X}_{..j}$.

This model is manifest in terms of measurement error, being based on a single indicator for the predictor and the outcome variable, but latent in terms of sampling error, assuming a latent aggregation of the level 1 variables to form the level 2 constructs. Therefore, it is called a "manifest latent" approach.

A.3.3 Model 3: The Latent Manifest Approach

Model 3 makes multivariate adjustments for measurement error in the observed by data viewing the outcome of the students at the different time points as latent constructs measured by multiple indicators. Typically, the available items that are used to measure individual-level variables can be used as different manifestations of the same underlying latent construct. For the purposes of the present investigation this approach was followed for self-concept measures. To create multiple indicators for the attainment measures, parcels were formed by averaging across different combinations of items. The latent factor at level two is measured by multiple indicators that are formed by taking the average of the multiple indicators at the student-level within each school (manifest aggregation). Again, I denote the number of indicators used for the baseline test by K and the number of indicators used for the outcome variable by L. Thus, if we assume that the number of individuals in cluster j is given by n_j, then level-2 multiple indicators are given by $\overline{X}_{k.j} = \frac{1}{n_j}\sum_{i=1}^{n_j} X_{kij}$

The measurement model at level 1 is given by:

$$X_{kij} - \overline{X}_{k.j} = \mu_{kx} + \lambda_{k,W}U_{xij} + R_{xkij}; k = 1,..., K \text{ (A.3.5)}$$

The measurement model at level 2 is:

$$\overline{X}_{k.j} = \mu_{kx} + \lambda_{k,B}U_{xj} + R_{xkj}; k = 1,..., K \text{ (A.3.6)}$$

The within and the between factor loadings are constrained to be equal cross the two levels (assumption of invariance at level 1 and level 2) so that $\lambda_{k,w} = \lambda_{k,B}$ for $k = 1,..., K$.

Each one of the observed indicators of the final attainment can be decomposed in a within and a between latent part as follows:

$$Y_{lij} = \mu_{ly} + \lambda_{ly,W} U_{yij} + R_{lyij} + \lambda_{ly,B} U_{yj} + R_{ylj}; l = 1,.., L. \quad \text{(A.3.7)}$$

The coefficients of the within part of the model (A.3.7), $\lambda_{ly,W}, l = 1,..., L$ represent the within factor loadings while $\lambda_{ly,B}, l = 1,..., L$ are the between-factor loadings, R_{lyij} are the residuals at level 1, and R_{lyj} are the residuals at Level 2. U_{yij} are the unobserved true scores at level 1 and U_{yj} are the unobserved scores at level 2.

The structural model that expresses the relationship between the student-level final achievement with the level 1 prior achievement and the level 2 prior achievement is:

$$U_{yij} = \beta_0 + \gamma_{10} U_{xij} + \gamma_{01} U_{xj} + \delta_{0j} + \varepsilon_{ij} \quad \text{(A.3.8)}$$

In this third approach, the school-level attainment is latent, being based on multiple indicators. Nevertheless, no adjustments are made for sampling error: The school-level indicators are given by manifest aggregation (by the observed school average score of the student-level indicators).

A.3.2 Model 4: The Doubly Latent Approach

The last model being attempted is doubly latent since it controls both for measurement error at level 1 and level 2 and for sampling error. The decomposition for response variable Y is the same as for Model 3 (see equation A.3.7).

The decomposition of the predictor X is as follows:

$$X_{kij} = \mu_{xk} + \lambda_{kx,W} U_{xij} + R_{xkij} + \lambda_{kx,W} U_{xj} + R_{xkj}; k = 1,.., K. \quad \text{(A.3.9)}$$

The structural model in this approach is given by:

$$U_{yij} = \beta_0 + \gamma_{10} U_{xij} + \gamma_{01} U_{xj} + \delta_{0j} + \varepsilon_{ij} \quad \text{(A.3.10)}$$

Note that the structural part of Model 3 (equation A.3.8) and Model 4 (equation A.3.10) are the same. The only difference between the latent manifest and the doubly latent approach is the way the level 2 indicators are constructed: In the former, manifest aggregation is assumed. This last approach forms the multiple indicators at level 2 assuming latent aggregation both for the predictor variable X (just like Model 2) and the response variable Y (as in Model 3). Therefore, it is "doubly latent", in that it controls both for measurement and for sampling error.

In all of my models I used group mean centring**.** In addition to the estimates of the model parameters, in my analysis I calculated an additional parameter, the difference between the between group coefficient γ_{01} and the within-group coefficient γ_{10}; this is an estimate of the compositional effect. In order to be able to do so, the measurement invariance of the latent factor structures (Marsh, Muthén, Asparouhov, Lüdtke, Robitzch, Morin and Trautwein, 2009,; Mehta and Neale, 2005; Meredith, 1993; Widaman and Reise, 1997) across the two levels (student – level 1 and school – level 2) was established. Mplus (Version 6) was used throughout my analysis. One advantage of the Mplus package in comparison with others that could be used for the purposes of my analysis is that it also provides a standard error for this estimate that allows for significance tests to be carried out.

A.4 Supplementary Analysis to Study 1a

A.4.1 Verification of the Negative Compositional Effect Using Different Approaches to Missing Data

In order to verify that the negative compositional effect which was detected in Study 1a was not a result of the way in which unit and item non-response was treated, I tried four distinct approaches that differed in (i) the way in which unit non-response was defined and (ii) the way in which item non-response was treated for units that were actually considered as non-missing in the analysis.

I remind the reader that missing units in the original analysis were identified as those units not having a matching assessment in the corresponding year as well as any case with five of fewer items attempted in the test (see "Defining unit non-response" under section 3.3.3 for Study 1a). Let me refer to this approach as the "more than five" rule. Moreover, item non-response was treated by replacing missing items with the probability of answering the item correctly (see "*Treatment of item non-response for mathematics achievement data"* under section 3.3.3 for Study 1a)

As a supplementary analysis to Study 1a, I considered the following modifications as to the way in which unit non-response and item non-response was defined: (i) Treatment of unit non-response in accordance to the "more than five" rule but treatment of item non-response by replacing missing items with zero (treating them simply as wrong). (ii) Treatment of unit non-response following the "more than one" rule. In this approach, I considered as non-missing units those with at least one item completed ("more than one" rule). The percentages of missing and non-missing cases in the dataset according to this rule are displayed among with the percentages of missing and non-missing cases in the dataset according to the "more than five" rule in Table 3.2 of Study 1a. There are no substantial differences in the number of students included in the analysis in each case. In this approach, item non-response was treated by replacing missing

items with the probability of answering the item correctly. (iii) In the last approach to missing data that I tried, I used the "more than one" rule to define missing cases and I replaced missing items with zero, i.e. I treated them as incorrect.

All three supplementary analyses were replications of the original study (see section 3.3.5 on the details of the statistical analysis for Study 1a). The results are displayed in Table A.2.The negative compositional effect was robust for all the distinct ways of treating unit and item non-response.

A.4.2 Verification of the Negative Compositional Effect Using Different Criteria for the Inclusion of Schools and Students in the Sample for the Analysis

In a different set of analyses, subsets of schools and students from the original database were incorporated and the analysis was replicated for each of these smaller datasets. I remind the reader that in the original analysis (see "Measures and Data Samples" section for Study 1a) I based my analysis on data from schools that participated in both year one and year four educational assessments - identified as those with at least one student with data on both measurement occasions (year one and year four). From these schools, I selected those students who had taken the PIPS test in year one or year four, even if they had missing data in one of those two years**.**

Subsamples of the dataset used for the original analysis were selected in the following ways: (i) By excluding all the students with missing data in year four and (ii) By selecting only those students who appeared to be in the same school at year three or year five as well instead of only at year one or at year four. My intention in both of these analyses was to obtain a more representative sample of the students in each school, using only the students that remained in the school for an adequate period of time. (iii) By removing from my initial sample all the schools that had fewer than ten students in year one or in year four. Thus, schools with small numbers of students were excluded from this analysis. The results (see Table A.3) were consistent for all different analyses and all revealed a negative compositional effect of school average mathematics ability on students' subsequent ability.

Table A. 2 Consistency of the results across different ways of the treatment of unit and item non-response

Models Tried	Original analysis (more than five items/use of a probability parameter)		More than five items/ Replacing missing items with zero		More than one item/ Use of a probability parameter		More than one item/ Replacing missing items with zero	
	$\beta_{within}\ (s.e.)$	$\beta_{com}\ (s.e.)$	$\beta_{within}\ (s.e.)$	$\beta_{com}\ (s.e.)$	$\beta_{within}\ (s.e.)$	$\beta_{com}\ (s.e.)$	$\beta_{within}\ (s.e.)$	$\beta_{com}\ (s.e.)$
	$ES_{\beta_{within}}\ (s.e.)$	$ES_{\beta_{com}}\ (s.e.)$	$ES_{\beta_{within}}\ (s.e.)$	$ES_{\beta_{com}}\ (s.e.)$	$ES_{\beta_{within}}\ (s.e.)$	$ES_{\beta_{com}}\ (s.e.)$	$ES_{\beta_{within}}\ (s.e.)$	$ES_{\beta_{com}}\ (s.e.)$
Model 1	.681 (.006	- .069 (.031)	.0678 (.007)	-.070 (.031)	.685 (.007)	-.066 (.031)	.641 (.008)	-.075 (.029)
	1.264 (.013)	*- .034 (.015)*	*1.263 (.014)*	*-.034 (.015)*	*1.266 (.013)*	*-.032 (.015)*	*1.179 (.015)*	*-.038 (.015)*
Model 2	.681 (.006)	- .077 (.037)	.678 (.007)	-.079 (.037)	.685 (.007)	-.074 (.037)	.614 (.008)	-.086 (.034)
	1.288 (.013)	*- .034 (.016)*	*1.287 (.014)*	*-.035 (.016)*	*1.290 (.013)*	*-.033(.016)*	*1.202 (.015)*	*-.04 (0.16)*
Model 3	.845 (.011)	- .129 (.036)	.833 (.011)	-.126 (.036)	.847 (.011)	-.126 (.035)	.752 (.011)	-.125 (.033)
	1.425 (.015)	*- .059 (.016)*	*1.416 (.015)*	*-.058 (.017)*	*11.425 (.015)*	*-.058 (.016)*	*1.338 (.017)*	*-.061 (.016)*
Model 4	.844 (.011)	- .144 (.041)	.833 (.011)	-.141 (.041)	.846 (.011)	-.141 (.041)	.752(.011)	-.142 (.038)
	1.452 (.015)	*-.0600 (.017)*	*1.443 (.015)*	*-.059 (.018)*	*1.451 (.015)*	*-.059 (.017)*	*1.363 (.017)*	*-.063 (.017)*

[1]*Notes*. Model 1 is the Doubly Manifest approach, Model 2 is the Manifest Latent approach, Model 3 is the Latent Manifest approach and Model 4 is the Doubly Latent approach. In the original analysis those students who answered less than five items were treated as missing cases. The probability of answering the item correct was used to replace the missing values for those students that had at least five items completed. In the second analysis, again students with less than five completed items were taken to be missing but the non-missing items for the others were assigned a value of zero, being considered as wrong. The third approach is analogous to the original analysis, but I only treated as missing cases the students with at least one item answered. In the same way, the fourth approach is analogous to the second approach. The parameter β_{within} is the within group effect, while β_{com} is the compositional effect. The abbreviation is used to denote the associated standard error while $ES_{\beta_{within}}$ and $ES_{\beta_{com}}$ are used to denote the effect size estimate for the within group effect and the between group effect respectively.

Table A. 3: Investigating whether the results found in the original analysis are consistent across different ways of selecting which schools and students from the sample to include in the analysis.

Models Tried	Original Analysis		Students with full data at year four		Students in the same School at year three or year five		Schools with less than ten students removed	
	$\beta_{within}\ (s.e.)$ $ES_{\beta_{within}}\ (s.e.)$	$\beta_{com}\ (s.e.)$ $ES_{\beta_{com}}\ (s.e.)$	$\beta_{within}\ (s.e.)$ $ES_{\beta_{within}}\ (s.e.)$	$\beta_{com}\ (s.e.)$ $ES_{\beta_{com}}\ (s.e.)$	$\beta_{within}\ (s.e.)$ $ES_{\beta_{within}}\ (s.e.)$	$\beta_{com}\ (s.e.)$ $ES_{\beta_{com}}\ (s.e.)$	$\beta_{within}\ (s.e.)$ $ES_{\beta_{within}}\ (s.e.)$	$\beta_{com}\ (s.e.)$ $ES_{\beta_{com}}\ (s.e.)$
Model 1	.681 (.006) *1.264 (.013)*	- .069 (.031) *- .034 (.015)*	.683 (.007) *1.266 (.014)*	- .077 (.032) *- .037 (.015)*	.683 (.007) *1.262 (.015)*	- .067 (.036) *- .032 (.017)*	.681 (.007) *1.268 (.013)*	-.076 (.032) *-.036 (.015)*
Model 2	.681 (.006) *1.288 (.013)*	- .077 (.037) *- .034 (.016)*	.682 (.007) *1.289 (.014)*	- .09 (.037) *- .04 (.017)*	.682 (.007) *1.285 (.015)*	- .076 (.042) *- .034 (.019)*	.681 (.007) *1.286 (.014)*	-.08 (.037) *-.035 (.017)*
Model 3	.845 (.011) *1.425 (.015)*	- .129 (.036) *- .059 (.016)*	.845 (.012) *1.427 (.015)*	- .139 (.036) *- .062 (.017)*	.840 (.012) *1.424 (.016)*	- .129 (.04) *- .058 (.018)*	.843 (.011) *1.429 (.015)*	-.136 (.036) *-.06 (.016)*
Model 4	.844 (.011) *1.452 (.015)*	- .144 (.041) *-.06 (.017)*	.845 (.012) *1.454 (.015)*	- .160 (.042) *- .066 (.018)*	.840 (.012) *1.450 (.017)*	- .145 (.047) *-.061 (.02)*	.842 (.011) *1.450 (.015)*	-.146 (.042) *-.061 (.018)*

[1]Note. . Model 1 is the Doubly Manifest approach, Model 2 is the Manifest Latent approach, Model 3 is the Latent Manifest approach and Model 4 is the Doubly Latent approach The parameter β_{within} is the within group effect, while β_{com} is the between group effect. The abbreviation is used to denote the associated standard error while $ES_{\beta_{within}}$ and $ES_{\beta_{com}}$ are used to denote the effect size estimate for the within group effect and the between group effect respectively. In the original analysis I selected all the schools that participated in both Year one and Year four assessment and subsequently, from each of these schools I chose those students who appeared to have been in the school and taken the assessment either in the Year one or Year four. The first modification that I made to my data (students with full data at Year four) was to select from the schools already selected in the original analysis, all the students who were in the school at Year four. Secondly (students in the same school at year three or year five), from the same set of schools, I selected those students who appeared to be in the school at year three or at year five as well as at Year one or at Year four. Lastly, I used the same data as the original analysis, having removed from my sample those schools than ten students. All these different analyses were performed in an attempt to achieve as representative sample of the population as possible.

Appendix B: The Regression Discontinuity Design in School Effectiveness Research

The regression discontinuity approach can be a valuable tool for assessing the effectiveness of the schools both in terms of equity and quality. Moreover, it can be used to investigate whether schools are differentially effective for distinct groups of students. Lastly, it can provide means to testingthe role of the school's composition (as well as of other school-level factors) in the schools' ability to promote knowledge. In what follows I explain how this can be achieved using mathematical equations.

B.1 Modelling the Absolute Effect of Schooling

In my analysis I fitted a series of nested models to model the absolute effect of schooling:

B.1.1 Equations for the Models Fitted to Investigate the Absolute Effect of Schooling and the Effect of Chronological Age on Students' Mathematics Achievement

I began with the empty model;

$$y_{ij} = \gamma_{00} + U_{0j} + \varepsilon_{ij} \quad \text{(B.1.1)}$$

I then control for the effect of age and the effect of schooling; the latter is the effect of a dummy variable that receives a value of one – when the individual is in the higher grade – and a value of zero – when the individual is in the lower grade. I refer to this as Model 2 in my analysis (in Model 1, I solely control for the effect of schooling):

$$y_{ij} = \gamma_{00} + \gamma_{10} age_{ij} + \gamma_{20} grade_{ij} + U_{0j} + \varepsilon_{ij} \quad \text{(B.1.2)}$$

To investigate relative differences across schools in their absolute effects I allowed the slope for the dummy variable for grade to be random across schools (Model 3 in my analysis):

$$y_{ij} = \gamma_{00} + \gamma_{10} age_{ij} + \gamma_{20} grade_{ij} + U_{0j} + U_{1j} grade_{ij} + \varepsilon_{ij} \quad \text{(B.1.3)}$$

Lastly, to investigate the extent to which there was a significant interaction between the effect of age and the effect of grade the significance of the effect of the product of the two variables was tested in my analysis:

$$y_{ij} = \gamma_{00} + \gamma_{10} age_{ij} + \gamma_{20} grade_{ij} + \gamma_{30} grade_{ij} age_{ij} + U_{0j} + U_{1j} grade_{ij} + \varepsilon_{ij} \quad \text{(B.1.4)}$$

In the equations given above y_{ij} is the achievement of student i in school j the variable age_{ij} is the age of the student centered around the cut-off age of nine years for primary schools and thirteen years for secondary school), γ_{ij} represents the effect of one year of chronological age, that is how much two students in the same grade but with one year's difference in their ages differ on average, in their achievement, γ_{2j} represents the effect of grade level, U_{0j} is the school-level residual that measures differences between schools in the achievement of their students in the lower grade, U_{1j} is the school-level residual that captures differences between schools in their absolute effect (of one year of extra schooling) and ε_{ij} is the student specific residual. The residuals U_{0j}, U_{1j} and ε_{ij} are assumed to be normally distributed with the school-level residuals (U_{0j}, U_{1j}) being independent of the student-level residual ε_{ij}.

The above model can be extended so that student-level variables are included as covariates, both as main effects and as interactions with the effect of grade.

B.1.2 Equations Fitted to Investigate the Effect of Adjustments for Student Background Variables on Added-Year Effects and to Investigate Differential School Effectiveness

The main effects of student-level background variables adjusted for in regression discontinuity models denote differences in the mean achievement of the students that differ by one unit in that student-level covariate, all other things being equal. The equation that describes a regression discontinuity model in which additional adjustments for $x_{3ij} \ldots x_{kij}$ student-level covariates are made with effects equal to $\gamma_{30} \ldots \gamma_{k0}$ is given extends equation (B.1.3) as follows:

$$y_{ij} = \gamma_{00} + \gamma_{10} age_{ij} + \gamma_{20} grade_{ij} + \gamma_{30} x_{3ij} + ... + \gamma_{k0} x_{kij} + U_{0j} + U_{1j} grade_{ij} + \varepsilon_{ij} \quad \text{(B.1.5)}$$

To investigate whether schools were differentially effective for distinct groups of students, the significance of the interactions between the effect of grade and the corresponding student-level characteristic, say x_{3ij} was tested in the following way:

$$y_{ij} = \gamma_{00} + \gamma_{10} age_{ij} + \gamma_{20} grade_{ij} + \gamma_{30} x_{3ij} + \gamma_{40} x_{3ij} grade_{ij} + U_{0j} + U_{1j} grade_{ij} + \varepsilon_{ij} \quad \text{(B.1.6)}$$

B.1.3 Equations Fitted to Examine the Relationships among Schools' Composition and Added-Year Effects

In order to test the effect γ_{21} of a school-level variable, say X on added-year effects, the following structural model was used:

$$y_{ij} = \gamma_{00} + \gamma_{10} age_{ij} + \gamma_{20} grade_{ij} + \gamma_{01} X + U_{0j} + (\gamma_{21} X + U_{1j}) grade_{ij} + \varepsilon_{ij} \quad \text{(B.1.7)}$$

In equation (B.1.7) γ_{01} denotes the main effect of the school-level variable; this reflects between school differences in the mean achievement of students who are the same age and are in the same grade.

B.2 Adjustments for Student-Level Background Variables in Regression Discontinuity Models

In Study 3, I investigated whether students with different background characteristics in terms of gender, ethnicity, socioeconomic status and home environment, were differentially influenced by schooling. To this end, the main effects of relevant student background variables and their interactions with grade were simultaneously tested for significance - one at a time. The results are displayed in Table B.1. The findings for primary school are displayed beside the findings for secondary school – the same set of student-level covariates were considered in both analyses. In what follows (secion B.2.1) I briefly discuss the findings on the main effects of the student background variables considered.

B.2.1 Main Effects of Student Background Variables

Gender: The positive effect of gender suggests that among the students who are in the same grade and are the same age, boys perform on average higher in mathematics than girls. It is interesting that both effects were the same size for primary school (β = .08, *se* = .036) as for secondary school (β = .08, *se* = .029).

Number of people in the student's house: The more siblings a student has, the lower his or her achievement will be. This is denoted by the negative effect of this school-level variable. Again the main effect is approximately the same both for primary (β =- .063, *se* = .012) and for secondary school data (β = -.065, *se* = .016).

Students' ethnicity: Four variables were available in TIMSS-95, related to students' ethnicity:

(i) A dummy variable denoting whether or not the student had been born in the UK (a value of one was taken if the student had not been born in the UK). The main effect of this variable was negative and significant; suggesting that students of the same age and in the same grade achieve, on average higher, if they had been born in England.

(ii) A variable denoting how often students speak English at home. This effect was negative, and, since this variable would take higher values for students that spoke

English less frequently, it suggests that children who are less exposed to the English language will perform generally lower, provided that they are in the same grade and that they are the same age.

(iii) A set of two distinct dummy variables, denoting whether or not the student's mother or father had been born in the UK. The main effects were negative for both of these variables; therefore a foreign mother or father suggested lower mathematics achievement, in general. It is interesting that the effect of both these variables was twice as large for secondary school data as for primary school data.

Home possessions: Four distinct dummy variables were used to assess home possessions (calculator, computer, study desk, dictionary). These variables would take a value of one to denote absence of the relevant possession at the student's home. The negative effects of all four variables for both primary and secondary school data suggest that not being equipped with the relevant possession negatively affected students' mathematical outcomes.

B.2.2 The Effect of Adjustments for Student Background on Added-Year Effects

In a separate model, all significant student-level main effects were included in the same model. Instead of the three dummy variables related to whether or not the student and/or his or her parents had not been born in England, ethnicity was described by one overall measure that combines the information provided by each of the three that were used. In the same way, home possession was also described by a single variable combining the information of the four dummy variables used in the initial exploratory analysis.

The results for the main effects of each of the covariates when considered simultaneously are displayed in Table B.2.

Table B. 1 Adjusting for the main effects and interactions of student-level background variables: Evidence for differential school effectiveness.

Student-level variable considered	Primary School data		Secondary School data	
	Main Effect	Interaction with grade	Main Effect	Interaction with grade
Gender	.08 (.036)	.023 (.048)	.08 (.029)	-.078 (.086)
Number of people in the student's house	-.063 (.012)	No interaction	-.065 (.016)	.006 (.021)
Number of books in the student's house	.160 (.013)	.016 (.02)	.214 (.016)	-.002 (.019)
Were you born in the UK?	-.340 (.066)	-.041 (.099)	-.258 (.094)	.031 (.142)
How often do you speak English at home?	-.364 (.05)	.051 (.084)	-.480 (.105)	.092 (.131)
Was your mother born in the UK?	-.05 (.051)	.043 (.066)	-.136 (.069)	.093 (.086)
Was your father born in the UK?	-.084 (.046)	.002 (.068)	-.195 (.062)	.053 (.084)
Do you have a calculator at home?	-.472 (.051)	-.103 (.082)	-.808 (.187)	.317 (.266)
Do you have a computer at home?	-.138 (.049)	.029 (.07)	-.037 (.061)	.031 (.082)
Do you have a study desk for your own use?	-.191 (.044)	-.03 (.064)	-.338 (.065)	.001 (.099)
Do you have a dictionary at home?	-.548 (.06)	-.058 (.095)	-.713 (.093)	.054 (.155)

Note. The effect of each of the background variables and the corresponding interaction term is tested in a separate model.

Table B. 2: Adjusting for the main effects of student-level background variables simultaneously

Student-level variable considered	Primary School data	Secondary School data
Gender	.103 (.026)	.028 (.028)
Number of people in the student's house	-.051 (.009)	-.059 (.011)
Number of books in the student's house	.133 (.012)	.192 (.013)
How often do you speak English at home?	-.235 (.049)	-.318 (.077)
Ethnicity	.009 (.02)	No significant effect
Home possessions	-.183 (.019)	-.127 (.028)

Note. The main effects for all of the variables listed in this table are examined simultaneously.

The most interesting finding of this second analysis which examines the simultaneous effects of student background is that, after adjustments for language spoken at home, the effect of ethnicity completely disappeared. This finding suggests that as long as the students use the language in which they are taught in the school in everyday communication with their family, then any potential negative effects of their not having been born in the UK disappear.

The main effects for the rest of the background variables considered followed a similar pattern as to when each of the covariates was considered individually (see Table B.1). A brief discussion of these findings ias given in what follows:

Gender: The effect of gender was positive and significant both for primary and secondary school data suggesting that boys on average outperformed girls who were the same age and in the same grade.

Number of people in house: The number of people in the students' house exerted a negative and significant effect on academic achievement and it was of approximately the same extent for both populations: The more children there were in a student's family, the lower the achieved results in mathematics.

Number of books in the house: More books in the house, a variable that measures to some extent the degree to which the family of the student fosters reading and learning, had a positive effect on achievement both for primary and secondary school data.

Frequency of English spoken at home: The frequency of English spoken at home had a negative and highly significant effect on achievement: Students who used English less frequently to communicate at home had on average lower achievement in mathematics.

Home possessions: A negative effect of possessing a computer/calculator/study desk/dictionary was found both with primary and secondary school data.

B.3 The Impact of Measurement Error in the Response Variable on Regression Discontinuity Estimates

Measurement error in the response variable in multiple regression models does not result in bias in the estimated effects of the covariates controlled for in the models, given that the corresponding independent variable is measured perfectly. It only leads to larger standard errors.

This can become clear to the reader, given that the formula for the standard error of the effect of a given covariate, X_k , in a multiple regression model that describes the relationship between a response variable Y and a set of K covariates, this set being denoted by H , is given by the formula (Williams, 2014):

$$s_{b_k} = \sqrt{\frac{1 - {R_{YH}}^2}{(1 - R^2_{X_k G_k})(N - K - 1)}} * \frac{s_Y}{s_X} \text{ (B.3.1)}$$

In equation (B.2.1) we have s_{b_k} being the standard error of the point estimate of the effect of the variable X_k , say b_k , R^2_{YH} is the multiple r-square obtained by regressing the response variable Y on the set of independent covariates H , $R^2_{X_k G_k}$ is the multiple r-square of regressing the covariate X_k on the set of all the X variables except X_k , N the number of observations in the data, K the number of covariates controlled for in the model, s_Y in the standard error for variable Y and s_X is the standard error for the variable X .

When measurement error is prevalent in the response variable, a downward bias is observed in R^2_{YH} while s_Y is larger due to the prevalence of random error in the response variable so that s_{b_k} is estimated larger.

Appendix C: The imprecise nature of value added assessment

The use of value added models for ranking educational institutions has been widely debated (Yang, Goldstein, Rath and Hill, 1999; Leckie and Goldstein, 2009; Goldstein and Spiengelhalter, 1996, Teddie and Reynolds, 2000). Here, I discuss some of the general criticisms surrounding the use of this approach for school assessment. In this way, I demonstrate how difficult it is to decide on the right methodology that should be followed to assess the performance of schools: Value added assessment is a very complicated issue, and researchers may never find the perfect way to model the actual estimates of the effectiveness of educational institutions.

One of the criticisms that pervades the use of league tables (Leckie and Goldstein, 2009) is that the information which is published each year is based on the current performance of a cohort of students who entered the secondary school a number of years earlier while, for choosing a school, it is the future performance of the current cohort that is of interest. As Gray, Goldstein and Thomas (2001) suggest: "Past performance is not necessarily a guide to future returns". When the uncertainty associated with the fact that we make predictions for the future is taken into account in our analysis, the derived confidence intervals associated with the value added estimate for each school become very wide - to the extent where any differences between the schools' effectiveness become non-significant (Leckie and Goldstein, 2009). This criticism relates to the issue of stability of school effects discussed in the section 6.4.6 of the discussion.

Indeed, it has been claimed (Marsh, Nagengast, Fletcher and Televantou, 2011) that the amount of variance explained by contextual value added models incorporated in educational effectiveness is very small (approximately 5%), so they may not be useful in distinguishing schools based on their performance in terms of improving the students' outcomes. Marsh, et al. (2011) suggest that a more appropriate analysis should take into account more levels of

variation in the students' outcomes, considering for instance, the effects of classes or teachers as well as the effects of schools (see section on teacher effects).

Other criticisms of the conventional value added modelling approach relate to the issue of differential effectiveness. They claim that the value added modelling masks the fact that some schools may be more effective for high performing pupils and less effective for low performing pupils. For example, in a relevant study, Dearden, Mickelwright and Vignoles, (2011) suggest that the ranking of the schools vary substantially for different prior attainment groups of pupils.

I have only listed some part of the criticisms that surround the use of (contextual) value added models to assess educational effectiveness. It should be clear by now that comparisons between educational institutions based on value added models and contextual value added models should be treated with care (Goldstein, Huiqqi, Rath and Hill, 2000). Value added estimation can only be a valuable tool for SER if it is used in the appropriate way: It should be best regarded as one indicator of the educational quality alongside others (Yang, Goldstein, Rath and Hill, 1999), and should not be the sole basis for high-stake decisions. For example, outlier institutions can be identified using value added modelling and then case studies can be performed in individual schools to investigate the reasons for their effectiveness or under-performance.

Despite the difficulties associated with the estimations of the value added of schools, the school league tables are here to stay. The task of educational researchers is to do the best that they can to provide accurate assessment of educational institutions.

Appendix D: MPLUS Setup Files

D.1 Testing the Role of Academic Self-Concept at the End of Year Four as a Mediator of the Negative Compositional Effect of School Average Achievement at the End of Year One on Individual Achievement at the End of Year four

```
TITLE: test of mediation using complex design;
 !In this file I fix the loading for the first indicators to
 !be equal to their values in the standardized solution
 !(STDYX standardization).In this way, I obtain standardized
     solutions
 !(factors with variance equal to one.

DATA: File is "mplus_data.dat";

TYPE IS IMPUTATION;

VARIABLE: Names are
 pup numAs Y1ass Y1sch Y1acy Y2ass Y2sch Y2acyr
 Y3ass Y3schl Y3acyr Y4ass Y4schl Y4acyr Y5ass
 Y5schl Y5acyr Y6ass Y6schl Y6acy schid A1Y1
 A2Y1 A3Y1 A4Y1 A5Y1 A1Y4 A2Y4 A3Y4 A4Y4 A5Y4
 Zpar11 Zpar12 Zpar13 Zpar41 Zpar42 Zpar43 Zpar44
 zach1 zach4 zach1m zach4m Zpar11m Zpar12m Zpar13m
 Zpar41m Zpar42m Zpar43m Zpar44m attm1 attm4
 A1Y1m A2Y1m A3Y1m A4Y1m A5Y1m
 A1Y4m A2Y4m A3Y4m A4Y4m A5Y4m;

 USEVAR ARE
  zpar11 zpar12 zpar13 zpar41 zpar42 zpar43 zpar44
  zpar11m zpar12m zpar13m A1Y4 A2Y4 A3Y4 A4Y4 A5Y4;
  cluster = schid;
  missing are all (999);
```

```
DEFINE:
 center zpar11-zpar13 (grandmean);

ANALYSIS: Type is Complex ;
 ESTIMATOR IS MLR;

MODEL:
 MACH1_W by zpar11@0.826      ;
 MACH1_W by zpar12;
 MACH1_W by zpar13;

 MACH4 by zpar41@0.863      ;
 MACH4 by zpar42;
 MACH4 by zpar43;
 MACH4 by zpar44;

 MASC4 by A1Y4@0.482;
 MASC4 by A2Y4;
 MASC4 by A3Y4;
 MASC4 by A4Y4;
 MASC4 by A5Y4;

 MACH1_B by zpar11m@0.938;
 MACH1_B by zpar12m;
 MACH1_B by zpar13m;

 MACH4 on MACH1_W (T4T1ACH);
 MACH4 on MASC4 (T4ACT4SC);
 MASC4 on MACH1_W (T4SCT1AC);
 MACH4 on MACH1_B (ContSC);
 MASC4 on MACH1_B (ContAC);
 MACH1_B on MACH1_W;
```

MODEL CONSTRAINT:

```
 new(bc_ind);
 bc_ind=ContSC*T4ACT4SC;
 new(bc_tot);
 bc_tot=ContAC+bc_ind;
```

OUTPUT: stand SVALUES;

D.2 Assessing the Magnitude of the Effects of School Composition Applying the Regression Discontinuity Approach in a Multilevel Structural Equation Modelling Framework

```
TITLE: The effect of school average achievement on school added-
year effects using multilevel structural equation models with
cross-level latent interaction (s on MACH_B);

DATA: File is "primary_timms95_data.dat";

VARIABLE: Names are

 idschl idstd itsex idgrd
 age tot hou sen mathL
 totm1 whom1 fapm1 gemm1 dapm1
 age_cr grade inter sex home
 book comm std1 home_m book_m
 gender gender_m G4_mean G5_mean ;
 USEVAR ARE
  whom1 fapm1 gemm1 dapm1
  age_cr grade hou ;
  within=grade age_cr ;
  cluster=IDSCHL;
  weight is hou;
  missing are all(999)
```

```
ANALYSIS: Type is two-level random; algorithm=integration;
integration=10;

MODEL:

 %within%

  MACH_W by whom1 (1);

  MACH_W by fapm1 (2);

  MACH_W by gemm1 (3);

  MACH_W by dapm1 (4);

  MACH_W on age_cr (b20);

  INT|grade XWITH MACH_B;

  MACH_W on grade INT;

 %between%

  MACH_B by whom1 (1);

  MACH_B by fapm1 (2);

  MACH_B by gemm1 (3);

  MACH_B by dapm1 (4);

  MACH_B (U0j);

OUTPUT: sampstat tech1;
```

Bibliography

Alwin, D. F. and Otto, L. B. (1977).High school context effects on aspirations. *Sociology of Education,* 50, pp. 259-273.

Alexander, K.L., Fennessey, J., McDill, E.L. and D'Amico, R.J. (1979) School SES influences--composition or context? *Sociology of Education,* 52 (4), pp. 222-237.

Ballou, D., Sanders, W. and Wright, P. (2004) Controlling for student background in value added assessment of teachers. *Journal of Educational and Behavioral Statistics,* 29 (1), pp. 37-65.

Bandura, A. (1986) *Social foundations of thought and action: A social cognitive theory.* Englewood Cliffs, NJ: Prentice-Hall.

Barr, R. and Dreeben, R. (1983) *How schools work.* Chicago: University of Chicago Press.

Barr, R. and Dreeben, R. (1991). Grouping students for reading instruction. In R. Barr, M. L. Kamil, P. B.Mosenthal and P. D. Pearson (Eds.), *Handbook of reading research, Vol. II* (pp. 885–910). New York: Longman.

Bedard, K. and Dhuey, E. (2006) The persistence of early childhood maturity: International evidence of long-run age effects. *The Quarterly Journal of Economics,* 121 (4), pp. 1437-1472.

Bentler, P. M. and Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin,* 88, pp. 588-606.

Bliese, P.D. (2000) Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In: K.J. Klein and S.W. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp.349 - 381). San Francisco: Jossey - Bass.

Bondi, L. (1991) Attainment at primary schools: an analysis of variations between schools. *British Educational Research Journal,* 17 (3), pp. 203-217.

Bourdieu, P. (1993) *Sociology in question.* Sage Publications Limited.

Boyd, L., and Iverson, G. (1979). Contextual analysis: Concepts and statistical techniques. Belmont, CA: Wadsworth.

Brennan, R. L. (2001) *Generalizability theory*. New York: Springer-Verlag.

Brennan, R.L. (2010) Generalizability theory and classical test theory, *Applied Measurement in Education*, 24 (1), pp. 1-21

Browne W. J. (2004) *MCMC estimation in MLwiN. Version 2.0*. London, Institute of Education, http://www.cmm.bristol.ac.uk/MLwiN/download/manuals.shtml

Browne, M.W.and Cudeck, R. (1993) Alternative ways of assessing model fit. In K.A. Bollen and J.S. Long (Eds.). *Testing Structural Equation Models* (pp. 136-163). London: Sage Publications.

Browne, W.J., Goldstein, H. and Rasbash, J. (2001) Multiple membership multiple classification (MMMC) models. *Statistical Modelling*, 1 (2), pp. 103.

Bryk, A.S. and Raudenbush, S.W. (1992) *Hierarchical linear models: Applications and data analysis methods.* Newbury Park, CA: Sage Publications.

Burstein, L. (1980) The analysis of multilevel data in educational research and evaluation. *Review of Research in Education*, 8 (1), pp. 158 - 233.

Byrne, B. M. and Shavelson, R.J. (1986) On the structure of adolescent self-concept. *Journal of Adolescent Psychology*, 78 (6), pp. 474-481.

Cahan, S. and Cohen, N. (1989) Age versus schooling effects on intelligence development. *Child Development*, 60 (5), pp. 1239-1249.

Cahan, S. and Davis, D. (1987) A between-grade-levels approach to the investigation of the absolute effects of schooling on achievement. *American Educational Research Journal*, 24 (1), pp. 1-12.

Callahan, R.E. (1962) *Education and the Cult of Efficiency.* Chicago: The University of Chicago Press.

Campbell, R.J., Kyriakides, L., Muijs, R.D., and Robinson, W. (2003). Differential teacher effectiveness: towards a model for research and teacher appraisal. *Oxford Review of Education,* 29 (3), pp. 347-362. Retrieved from http://www.jstor.org/stable/3595446

Ceci (1991) How much does schooling influence general intelligence and its cognitive components? A reassessment of the evidence. *Developmental Psychology,* 27 (5), pp. 703-722.

Chanal, J., Marsh, H.W., Sarrazin, P. and Bois, J. (2005) Big-fish-little-pond effects on gymnastics self-concept: Social comparison processes in a physical setting. *Journal of Sport and Exercise Psychology,* 27 (1), pp. 53-70.

Chapman, J.W. and Tunmer, W.E. (1995) Development of young children's reading self-concepts: An examination of emerging subcomponents and their relationship with reading achievement. *Journal of Educational Psychology,* 87 (1), pp. 154.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling,* 14 (3), pp. 464-504.

Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA teststatistic in structural equation models. *Sociological Methods and Research,* 36, pp. 462– 494.

Cheung, G.W. and Rensvold, R.B. (2001). The effects of model parsimony and sampling error on the fit of structural equation models. *Organizational Research Methods,* 4, pp. 236-264.

Cheung, G. W. and Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling,* 9, pp. 233–255.

Cohen, J. (1998) *Statistical power analysis or the behavioural sciences.* (2nd ed.) New York: Academic Press.

Cohen, J. (1992) A power primer. *Psychologcal Bulletin.* 112, (1), pp. 155-159.

Cliffordson, C. (2010) Methodological issues in investigations of the relative effects of schooling and age on school performance: The between-grade regression discontinuity

design applied to Swedish TIMSS 1995 data. *Educational Research and Evaluation,* 16 (1), pp. 39-52.

Cliffordson, C. and Gustafson, J. (2011, August). *Using instrumental variable regression techniques with fuzzy between-grade regression discontinuity designs to estimate effects of schooling and grade on achievement.* Paper presented in the symposium Causal Effects in Educational Research at the 14th Biennial EARLI Conference for Research on Learning and Instruction, Exeter, England.

Coe, R. and Fitz-Gibbon, C.T. (1998) School effectiveness research: Criticisms and recommendations. *Oxford Review of Education.* 24 (4), pp. 421-438.

Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J., Mood, A.M., Weinfeld, F.D. and Robert, L. (1966) *Equality of educational opportunity.* Washington, DC: US Government Printing Office.

Collins, L.M., Schafer, J.L. and Kam, C.M. (2001) A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods,* 6 (4), pp. 330-351.

Cook, T.D. (2008) "Waiting for life to arrive": a history of the regression-discontinuity design in Psychology, Statistics and Econometrics. *Journal of Econometrics*, 142, pp. 636-654.

Cook, T.D., Campbell, D.T. and Day, A. (1979) *Quasi-experimentation: Design & analysis issues for field settings.* Boston: Houghton Mifflin.

Coopersmith, S. (1967) *The antecedents of self-esteem.* San Francisco: Freeman

Crawford, C., Dearden, L. and Greaves, E. (2011) *Does when you are born matter? The impact of month of birth on children's cognitive and non-cognitive skills in England* (Report for the Nuffield Foundation ISBN: 978-1-903274-87-3). Retrieved from the Institute for Fiscal Studies website: http://www.ifs.org.uk/bns/bn122.pdf

Creemers, B.P.M.,and Kyriakides, L. (2010) School factors explaining achievement on cognitive and affective outcomes: Establishing a dynamic model of educational effectiveness. *Scandinavian Journal of Educational Research*, 54 (1), pp. 263-294.

Creemers, B.P.M. and Kyriakides, L. (2008) *The dynamics of educational effectiveness: a contribution to policy, practice and theory in contemporary schools*. London: Routledge.

Creemers, B.P.M., Kyriakides, L. and Sammons, P. (2010) *Methodological advances in educational effectiveness research.* London: Routledge.

Cronbach, L. J. (1984) Essentials of psychological testing, 4th ed., New York: Harper & Row.

Cronbach, L. J., Gleser, G. C., Nanda, H., and Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Cronbach, J., Linn, R. L., Brennan, R. L. and Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement*, 57, pp. 373–399.

Croon, M. A., and Van Veldhoven, M. J. P. M. (2007) Predicting group-level variables from variables measured at the individual-level: A latent variable multilevel model. *Psychological Methods*, 12, pp. 45–57.

Crone, D.A. and Whitehurst, G.J. (1999) Age and schooling effects on emergent literacy and early reading skills. *Journal of Educational Psychology*, 91 (4), pp. 604-614.

Dale, R. (1977) Educational markets and school choice. *British Journal of Sociology of Education*, 18, pp. 451-468.

Darlington, R.B., Royce, J.M., Snipper, A.S., Murray, H.W. and Lazar, I. (1980) Preschool programs and later school competence of children from low-income families. *Science*, 208 (440), pp. 202-204.

Davis, J.A. (1966) The campus as a frog pond: An application of theory of relative deprivation to career decisions for college men. *American Journal of Sociology*, 72, pp. 17-31.

De Fraine, B., Van Damme, J., Van Landeghem, G., Opdenakker, M. K. and Onghena, P. (2003). The effects of schools and classes on language achievement. *British*

Educational Research Journal, 29 (6), pp. 841–859. doi: 10.1080/0141192032000137330

Dearden, L., Micklewright, J. and Vignoles, A. (2011) The effectiveness of english secondary schools for pupils of different ability levels. *Fiscal Studies*, 32 (2), pp. 225-244.

Dearden, L. and Vignoles, A. (2011) Schools, markets and league tables. *Fiscal Studies*, 32 (2), pp. 179-186.

Degracie, J.S. and Fuller, W.A. (1972) Estimation of the slope and analysis of covariance when concominant variable is measured with error. *Journal of the American Statistical Association.* 67 (340), pp.930-937.

Dumay, X. and Dupriez, V. (2008) Does the school composition effect matter? Evidence from Belgian data. *British Journal of Educational Studies*, 56 (4), pp. 440-477.

Duru-Bellat, M., Le Bastard-Landrier, S., Piquee, C., and Suchaut, B. (2004). Social school mix and the experience of high school and primary school pupils. *Revue française de Sociologie*, 45 (3), pp. 441–468. Retrieved from http://www.cairn.info/revue-francaise-de-sociologie-2004-3-page-441.htm

Ferrão, M.E. and Goldstein, H. (2009) Adjusting for measurement error in the value added model: evidence from Portugal. *Quality and Quantity*, 43 (6), pp. 951-963.

Festinger, L. (1954) A theory of social comparison processes. *Human Relations*, 7, pp. 117–140.

Fielding, A., and Goldstein, H. (2006). *Cross-classified and multiple membership structures in multilevel models: An introduction and review.* Research Report RR791. Birmingham, UK: Department for Education and Skills, University of Birmingham.http://www.socscistaff.bham.ac.uk/fielding/Cross_classified_review_RR791.pdf

Fletcher, J. (2012). *Mismeasurement, mis-modelling and the estimation of effectiveness in education*. PhD Thesis, University of Oxford.

Foy, P. and Olson, J. F. (2009). *TIMMS 2007 user guide for the international database.* Boston: International Association for the Education of Evaluation Achievement.

Fuller, W. A. (1987). *Measurement error models.* London and New York: Wiley.

Gibbons, S., Machin, S. and Silva, O. (2008) Choice, competition, and pupil achievement. *Journal of the European Economic Association,* 6 (4), pp. 912-947.

Goe, L., Bell, C., Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. National Comprehensive Center for Teacher Quality, Washington DC.

Goldberger, A.S. (2008) Selection bias in evaluating treatment effects: Some formal illustrations. *Advances in Econometrics,* 21, pp. 1-31.

Goldstein, H. (2011) *Multilevel statistical models.* (2nd ed.). UK: Wiley.

Goldstein, H. (1995) *Multilevel statistical models.* (4th ed.). London: Edward Arnold.

Goldstein H. (1987) Multilevel covariance component models. *Biometrika.* 74 (2), pp.430-431

Goldstein, H., Burgess, S. and McConnell, B. (2007) Modelling the effect of pupil mobility on school differences in educational achievement. *Journal of the Royal Statistical Society-Series A,* 170 (4), pp. 941-954.

Goldstein, H., Huiqi, P., Rath, T. and Hill, N. (2000) *The use of value added information in judging school performance.* London, Institute of Education.

Goldstein, H., Kounali, D. and Robinson, A. (2008) Modelling measurement errors and category misclassifications in multilevel models. *Statistical Modelling,* 8 (3), pp. 243-261.

Goldstein, H. and Sammons, P. (1997) The influence of secondary and junior schools on sixteen year examination performance: a cross-classified multilevel analysis. *School Effectiveness and School Improvement*, 8 (2), pp. 219-230.

Goldstein, H. and Spiegelhalter, D.J. (1996) League tables and their limitations: statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society.Series A (Statistics in Society)*, 159 (3), pp. 385-443.

Gorard, S. (2010). Serious doubts about school effectiveness, *British Educational Research Journal,* 36 (5), 745–766.

Gray, J., Goldstein, H. and Thomas, S. (2001) Predicting the future: the role of past performance in determining trends in institutional effectiveness at A level. *British Educational Research Journal,* 27 (4), pp. 391-405.

Gray, J., Jesson, D. and Sime, N. (1990) Estimating differences in the examination performances of secondary schools in six LEAs: a multi-level approach to school effectiveness. *Oxford Review of Education,* 16 (2), pp. 137-158.

Guay, F., Larose, S. and Boivin, M. (2004) Academic self-concept and educational attainment level: A ten-year longitudinal study. *Self and Identity,* 3 (1), pp. 53-68.

Guay, F., Marsh, H.W. and Boivin, M. (2003) Academic self-concept and academic achievement: Developmental perspectives on their causal ordering. *Journal of Educational Psychology,* 95 (1), pp. 124-136.

Guldemond, H. and Bosker, R. (2009) School effects on students' progress-a dynamic perspective. *School Effectiveness and School Improvement,* 20 (2), pp. 255-268.

Hahn, J., Todd, P. and Van der Klaauw, W. (2008) Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica,* 69 (1), pp. 201-209.

Harker, R. (2004) *Compositional effects in school effectiveness studies: a New Zealand case study*. Paper presented at the AERA Annual Conference, San Diego, CA, USA, April. unpublished.

Harker, R. and Nash,R . (1996) Academic outcomes and school effectiveness: type 'A' and type 'B' effects, *New Zealand Journal of Educational Studies*, 32, pp. 143-170.

Harker, R. and Tymms, P. (2004) The effects of student composition on school outcomes. *School Effectiveness and School Improvement,* 15 (2), pp. 177-199.

Harris, D. N. and McCaffrey, D. F. (2010). Value added: Assessing teachers' contributions to student achievement. In M. Kennedy (Ed.), *Handbook of teacher assessment and teacher quality.* (pp. 251-282). San Francisco: Jossey Bass.

Hattie, J. (2009) *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.* Oxford: Routledge.

Hattie, J.A.C. (2002) Classroom composition and peer effects. *International Journal of Educational Research,* 37 (5), pp. 449-481.

Heck, R.H. and Moriyama, K. (2010) Examining relationships among elementary schools' contexts, leadership, instructional practices and added-year outcomes: a regression discontinuity approach. *School Effectiveness and School Improvement*, 21 (4), pp.377-408.

Helson, H. (1964) *Adaptation-level theory.* New York: Harper & Row.

Hess, R.D., Holloway, S.D., Dickson, W.P. and Price, G.G. (1984) Maternal variables as predictors of children's school readiness and later achievement in vocabulary and mathematics in sixth grade. *Child Development*, 55 (5), pp. 1902-1912.

Heyns, B. (1978) *Summer learning and the effects of schooling.* New York: Academic Press.

Hu, L. and 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), pp. 1-55.

Hu, L. and Bentler, P.M. (1998) Fit indices in covariance structure modeling: Sensitivity to under-parameterized model misspecification. *Psychological Methods,* 3 (4), pp. 424-453.

Husen, T. (1970). The effect of school structure upon utilization of ability: The case of Sweden and some international comparisons. In D. Swift (Ed.), *Basic readings in the sociology of education* (pp. 121–135). London: Routledge and Kegan Paul.

Hutchison, D. (2007) When is a compositional effect not a compositional effect? *Quality and Quantity,* 41 (2), pp. 219-232.

Hutchison, D. (2004) The effect of measurement errors on apparent group-level effects in educational progress. *Quality and Quantity,* 38 (4), pp. 407-424.

Hutchison, D. (1993) School effectiveness studies using administrative data. *Educational Research*, 35 (1), pp. 27-47.

Imbens, G.W. and Lemieux, T. (2008) Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142, pp.615-635.

Iverson, G. R. (1991) *Contextual analysis* (Sage University Paper Series on Quantitative Applications in the Social Sciences No.07–081). Newbury Park, CA: Sage.

Jaccard, J. and Wan, C.K. (1996) *LISREL approaches to interaction effects in multiple regression.* London: Sage Publications, Incorporated.

Jencks, C., Smith, M., Acland, H., Bane, M.J., Cohen, D., Gintis, H. and Heyns, B. and Michelson, S. (1972) *Inequality: A reassessment of the effects of family and schooling in America*. New York: Basic Books.

Joreskog KG (1970) A general method for analysis of covariance structures. *Biometrika*, 57, pp. 239-251.

Joreskog, K. G., and Sorbom, D. (1993). *LISREL* 8: *Structural equation modeling wirh the SIMPLIS* command language. Chicago: Scientific Software International.

Judd, C. M., and Kennedy, D. A. (1981). *Estimating the effects of social interventions.* New York:C ambridgeU niversityP ress.

Keeves, J.P., Hungi, N. and Afrassa, T. (2005) Measuring value added effects across schools: Should schools be compared in performance? *Studies in Educational Evaluation*, 31 (2-3), pp. 247-266.

Kish, L. (1965). *Survey sampling*. New York, NY: Wiley.

Kreft, I. G. G., de Leeuw, J., and Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research*, 30, pp. 1–21.

Kyriakides, L. (2008) Testing the validity of the comprehensive model of educational effectiveness: a step towards the development of a dynamic model of effectiveness. *School Effectiveness and School Improvement*, 19 (4), pp. 429-446.

Kyriakides, L., and Creemers, B. P. M. (2008). A longitudinal study on the stability over time of school and teacher effects on student learning outcomes. *Oxford Review of Education, 34*, 521–545.

Kyriakides, L., and Creemers, B.P.M. (2011) Can Schools Achieve Both Quality and Equity? Investigating the Two Dimensions of Educational Effectiveness. *Journal of Education for Students Placed at Risk*, 16 (4), pp. 237-254.

Kyriakides, L., Creemers, B.P.M., Teddlie, C. and Muijs, D. (2010). The International system for teacher observation and feedback: A theoretical framework for developing international instruments. In P. Peterson, E. Baker and B. McGaw (Eds.), *The international system for teacher observation and feedback: A theoretical framework for developing international instruments. Volume 3* (pp. 726-734). Oxford: Elsevier.

Kyriakides, L. and Luyten, H. (2009) The contribution of schooling to the cognitive development of secondary education students in Cyprus: an application of regression discontinuity with multiple cut-off points. *School Effectiveness and School Improvement*, 20 (2), pp. 167-186.

Kyriakides, L. and Tsangaridou, N. (2008). Towards the development of generic and differentiated models of educational effectiveness: a study on school and teacher Effectiveness in Physical Education. *British Educational Research Journal*, 34 (6), pp. 807-838.

Lauder, H., Hughes, D., Watson, S., Waslander, S., Thrupp, M., Strathdee, R., Simiyu, I., Dupuis, A., McGlinn, J. and Hamlin, J. (1999). *Trading in futures: Why markets in education don't work*, Open University Press: Buckingham.

Lazar, I., Darlington, R., Murray, H., Royce, J., Snipper, A. and Ramey, C.T. (1982) Lasting effects of early education: A report from the consortium for longitudinal studies. *Monographs of the Society for Research in Child Development*, 47 (2/3, Lasting Effects of Early Education: A Report from the Consortium for Longitudinal Studies), pp. i-151.

Lauder, H., Kounali, D., Robinson, T., and Goldstein, H. (2010) Pupil composition and account-ability: An analysis in English primary schools. *International Journal of Educational Research*, 49, pp. 49–68.

Lauder, H., Kounali, D., Robinson, T., Goldstein, H. and Thrupp, M. (2007) *Social class, pupil composition, pupil progress and school performance: an analysis of primary schools*. Retrieved from University of Bath website: http://www.bath.ac.uk/research/harps/Resources/The%20Effects%20of%20Pupil%20Composition%20in%20Primary%20Schools%20wbl.pdf

Leckie, G. (2009) The complexity of school and neighbourhood effects and movements of pupils on school differences in models of educational achievement. *Journal of the Royal Statistical Society: Series A*, 172 (3), pp.537-554.

Leckie, G. and Goldstein, H. (2009) The limitations of using school league tables to inform school choice. *Journal of the Royal Statistical Society, Series A*, 172 (1), pp. 835-872.

Liem, G. A. D., Marsh, H. W., Martin, A. J., McInerney, D. M. and Yeung, A. A. (2013). The big-fish-little-pond effect and a national policy of within-school ability streaming: Alternative frames of reference. *American Educational Research Journal*. 50 (2), pp. 326-370.

Little, T.D., Cunningham, W.A., Shahar, G. and Widaman, K.F. (2002) To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling*, 9 (2), pp. 151-173.

Lord, F.M., Novick, M.R. and Birnbaum, A. (1968) *Statistical theories of mental test scores*. Oxford, England: Addison-Wesley.

Lüdtke, O., Köller, O., Marsh, H.W., Urich, T. (2005) Teacher frame of reference and the big-fish-little-pond-effect. *Contemporary Educational Psychology*, 30 (3), pp. 263-285.

Lüdtke, O., Marsh, H.W., Robitzsch, A., Trautwein, U., Asparouhov, T. and Muthén, B. (2008) The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, 13 (3), pp. 203-229.

Lüdtke, O., Marsh, H.W., Robitzsch, A. and Trautwein, U. (2011) A 2× 2 taxonomy of multilevel latent contextual models: Accuracy–bias trade-offs in full and partial error correction models. *Psychological Methods*, 16 (4), pp. 444-467.

Lüdtke, O., Robitzsch, A., Thoemmes, F. and Trautwein, U. (2013) *Multilevel data do not always imply multilevel analysis: A comment on preacher, zyphur, and zhang (2010)*. Manuscript submitted for publication.

Luyten, H. (2006) An empirical assessment of the absolute effect of schooling: regression-discontinuity applied to TIMSS-95. *Oxford Review of Education*, 32 (3), pp. 397-429.

Luyten, H., Peschar, J. and Coe, R. (2008) Effects of schooling on reading performance, reading engagement, and reading activities of 15-year-olds in England. *American Educational Research Journal*, 45 (2), pp. 319-342.

Luyten, H., Tymms, P. and Jones, P. (2009) Assessing school effects without controlling for prior achievement? *School Effectiveness and School Improvement*, 20 (2), pp. 145-165.

Luyten, H. and Veldkamp, B. (2011) Assessing Effects of Schooling with cross-sectional data: between-grades differences addressed as a selection-bias problem. *Journal of Research on Educational Effectiveness*, 4 (3), pp. 264-288.

Marcoullides, G. A. and Kyriakides, L. (2010). Using generalizability theory. In B. P. M. Creemers, L. Kyriakides and P. Sammons (Eds.), *Methodological advances in Educational Effectiveness Research* (pp. 219–246). London: Routledge.

Marsh, H.W. (1974) *Judgmental anchoring: Stimulus and response variables.* Unpublished doctoral dissertation, University of California, Los Angeles.

Marsh, H. W. (1984a) Self-concept: The application of a frame of reference model to explain paradoxical results. *Australian Journal of Education*, 28, 165–181.

Marsh, H. W. (1984b) Self-concept, social comparison, and ability grouping: A reply to Kulik and Kulik. *American Educational Research Journal*, 21 (4), pp. 799–806.

Marsh, H.W. (1987) The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology*, 79 (3), pp. 280-295.

Marsh, H.W. (1990) The structure of academic self-concept: The Marsh/Shavelson model. *Journal of Educational Psychology,* 82 (4), pp. 623-636.

Marsh, H.W. (1991) Failure of high-ability high schools to deliver academic benefits commensurate with their students' ability levels. *American Educational Research Journal,* 28 (2), pp. 445-480.

Marsh, H.W. (1992) Content specificity of relations between academic achievement and academic self-concept. *Journal of Educational Psychology,* 84 (1), pp. 35-42.

Marsh, H. W. (1994). Using the national educational longitudinal study of 1988 to evaluate theoretical models of self-concept: The self-description questionnaire. *Journal of Educational Psychology,* 86, pp. 439–456.

Marsh, H. W. (2007a). Application of confirmatory factor analysis and structural equation modeling in sport/exercise psychology. In G. Tenenbaum & R. C. Eklund (Eds.), *Handbook of on sport psychology* (3rd ed.). New York: Wiley.

Marsh, H. W. (2007b). *Self-concept theory, measurement and research into practice: The role of self-concept in educational psychology.* Leicester: British Psychological Society

Marsh, H.W., Balla, J. R., and Hau, K. T. (1996). An evaluation of incremental fit indices: A clarification of mathematical and empirical processes. In G. A. Marcoulides and R. E. Schumacker (Eds.), *Advanced structural equation modeling techniques* (pp. 315–353). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Marsh, H.W., Balla, J.R. and McDonald, R.P. (1988) Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin,* 103 (3), pp. 391-410.

Marsh, H.W., Byrne, B.M. and Shavelson, R.J. (1988) A multifaceted academic self-concept: Its hierarchical structure and its relation to academic achievement. *Journal of Educational Psychology,* 80 (3), pp. 366-380.

Marsh, H.W., Byrne, B.M. and Yeung, A.S. (1999) Causal ordering of academic self-concept and achievement: Reanalysis of a pioneering study and revised reccomendations. *Educational Psychologist,* 34 (3), pp. 155-167.

Marsh.H.W., Chanal,J.P.and Sarrazin, P.G. (2006) Self-belief does not make a difference: A reciprocal effects model of the causal ordering of physical self-concept and gymnastics performance. *Journal of Sports Science.* 24 (1), pp.101-111.

Marsh, H.W., Chessor, D., Craven, R. and Roche, L. (1995) The effects of gifted and talented programs on academic self-concept: The big fish strikes again. *American Educational Research Journal*, 32 (2), pp. 285-319.

Marsh, H.W. and Craven, R.G. (2002). The Pivotal Role of Frames of Reference in Academic Self-Concept Formation: The" Big Fish-Little Pond" Effect. In F. Pajares and T. Urdan (Eds.), *Academic Motivation of Adolescents. Adolescence and Education Series* (pp. 83-124). US: Information Age Publishing Inc.

Marsh, H.W. and Craven, R.G. (2006) Reciprocal effects of self-concept and performance from a multidimensional perspective: Beyond seductive pleasure and unidimensional perspectives. *Perspectives on Psychological Science*, 1 (2), pp. 133-163.

Marsh, H.W., Craven, R. and Debus, R. (1998) Structure, stability, and development of young children's self-concepts: A multicohort–multioccasion study. *Child Development*, 69 (4), pp. 1030-1053.

Marsh, H.W. and Hau, K.T. (2007) Applications of latent-variable models in educational psychology: The need for methodological-substantive synergies. *Contemporary Educational Psychology,* 32 (1), pp. 151-170.

Marsh, H.W. and Hau, K.T. (2003) Big-Fish--Little-Pond effect on academic self-concept: A cross-cultural (26-country) test of the negative effects of academically selective schools. *American Psychologist,* 58 (5), pp. 364-376.

Marsh, H.W., Hau, K.T. and Wen, Z. (2004) In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*, 11 (3), pp. 320-341.

Marsh, H.W., Köller, O. and Baumert, J. (2001) Reunification of East and West German school systems: Longitudinal multilevel modeling study of the big-fish-little-pond effect on academic self-concept. *American Educational Research Journal*, 38 (2), pp. 321-350.

Marsh, H.W., Kong, C.K. and Hau, K.T. (2000) Longitudinal multilevel models of the big-fish-little-pond effect on academic self-concept: Counterbalancing contrast and reflected-glory effects in Hong Kong schools. *Journal of Personality and Social Psychology*, 78 (2), pp. 337-349.

Marsh, H.W., Lüdtke, O., Nagengast, B. , Morin, A.J.S., and von Davier, M. (2013). Why item parcels are (almost) never appropriate: Two wrongs do not make a right—camouflaging misspecification with item parcels in CFA models. *Psychological Methods*, 18, 257-284.

Marsh, H.W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A.J.S., Abduljabbar, A.S. and Köller, O. (2012) Classroom climate and contextual effects: Conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist*, 47 (2), pp. 106-124.

Marsh, H.W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B. and Nagengast, B. (2009) Doubly-latent models of school contextual effects: Integrating Multilevel and structural equation approaches to control measurement and sampling error. *Multivariate Behavioral Research*, 44 (6), pp. 764-802.

Marsh, H.W. and MacDonald Holmes, I.W. (1990) Multidimensional self-concepts: Construct validation of responses by children. *American Educational Research Journal*, 27, pp. 89-117.

Marsh, H.W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A.J.S. and Trautwein, U. (2009) Exploratory structural equation modeling, integrating CFA and EFA: application to students' evaluations of university teaching. *Structural Equation Modeling: A Multidisciplinary Journal*, 16 (3), pp. 439-476.

Marsh, H. W, Nagengast, B., Fletcher, J. and Televantou, I. (2011). Assessing educational effectiveness: policy implications from diverse areas of research. *Fiscal Studies*, 32 (2), pp. 279–295.

Marsh, H.W. and O'Mara, A.J. (2010) Long-term total negative effects of school-average ability on diverse educational outcomes. *Zeitschrift für Pädagogische Psychologie*, 24 (1), pp. 51-72.

Marsh, H.W. and Parker, J.W. (1984) Determinants of student self-concept: Is it better to be a relatively large fish in a small pond even if you don't learn to swim as well? *Journal of Personality and Social Psychology,* 47 (1), pp. 213-231.

Marsh, H.W. and Rowe, K.J. (1996) The negative effects of school-average ability on academic self-concepts: An application of multilevel modelling. *Australian Journal of Education,* 40 (1), pp. 65-87.

Marsh, H.W., Seaton, M., Kuyper, H., Dumas, F., Huguet, P., Régner, I., Buunk, A.P., Monteil, J.M. and Gibbons, F.X. (2010) Phantom behavioral assimilation effects: Systematic biases in social comparison choice studies. *Journal of Personality,* 78 (2), pp. 671-710.

Marsh, H.W., Seaton, M., Trautwein, U., Lüdtke, O., Hau, K.T., O'Mara, A.J. and Craven, R.G. (2008) The big-fish–little-pond-effect stands up to critical scrutiny: Implications for theory, methodology, and future research. *Educational Psychology Review,* 20 (3), pp. 319-350.

Marsh, H.W. and Shavelson, R. (1985) Self-concept: Its multifaceted, hierarchical structure. *Educational Psychologist*, 20 (3), pp. 107-123.

Marsh, H.W., Trautwein, U., Lüdtke, O., Köller, O. and Baumert, J. (2005) ASC, Interest, grades and standardised test scores: Reciprocal effects models of causal ordering. *Child Development*, 7, pp. 297-416.

Marsh, H.W., Trautwein, U., Lüdtke, O., Köller, O. and Baumert, J. (2006) Integration of multidimensional self-concept and core personality constructs: Construct validation and relations to well-being and achievement. *Journal of Personality,* 74 (2), pp. 403-456.

Marsh, H.W., Trautwein, U., Lüdtke, O., Baumert, J. and Köller, O. (2007) The big-fish-little-pond effect: Persistent negative effects of selective high schools on self-concept after graduation. *American Educational Research Journal,* 44 (3), pp. 631-669.

Marsh, H.W. and Yeung, A.S. (1997) Causal effects of academic self-concept on academic achievement: Structural equation models of longitudinal data. *Journal of Educational Psychology,* 89 (1), pp. 41-54.

Marx, R.W. and Winne, P.H. (1978) Construct interpretations of three self-concept inventories. *American Educational Research Journal, 15* (1), pp. 99-109.

Mehta, P.D. and Neale, M.C. (2005) People are variables too: multilevel structural equations modeling. *Psychological Methods,* 10 (3), pp. 259-284.

Meredith, W. (1993) Measurement invariance, factor analysis and factorial invariance. *Psychometrika,* 58 (4), pp. 525-543.

Mortimore, P., Sammons, P., Stoll, L., Lewis, D. and Ecob, R. (1988) *School matters: The junior years.* London: Paul Chapman.

Muijs, D. (2006) Measuring teacher effectiveness: Some methodological reflections. *Educational Research and Evaluation,* 12 (1), pp. 53-74.

Muijs, D., Kelly, T., Sammons, P., Reynolds, D. and Chapman, C. (2011) The value of educational effectiveness research – a response to recent criticism. *Research Intelligence*, 114, pp.24-25.

Muthén, L. K., and Muthén, B. O. (1998–2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Authors.

Nagengast, B. and Marsh, H.W. (2011) The negative effect of school-average ability on science self-concept in the UK, the UK countries and the world: the Big-Fish-Little-Pond-Effect for PISA 2006. *Educational Psychology,* 31 (5), pp. 629-656.

Nagengast, B., and Marsh, H.W. (2012) Big fish in little ponds aspire more: Mediation and cross-cultural generalizability of school-average ability effects on self-concept and career aspirations in science. *Journal of Educational Psychology,* 104, pp. 1033-1053.

Nash, R. (2003). Is the school composition effect real?: A discussion with evidence from the UK PISA Data. *School Effectiveness and School Improvement*, (14), pp. 441–457.

OECD (2008) *Measuring Improvements in Learning Outcomes: Best Practices to Assess the Value added of the Schools.* [e-book] Retrieved from http://www.oecd.org/document/54/0,3343,en_2649_39263231_41701046_1_1_1_37455,00.html [Accessed 5 May2009]

Opdenakker, M.C. and Van Damme, J. (2000) Effects of schools, teaching staff and classes on achievement and well-being in secondary education: Similarities and differences between school outcomes. *School Effectiveness and School Improvement,* 11 (2), pp. 165-196.

Osmond, R. (2000) The importance of mathematics to employers. *Teaching Mathematics Applications*, 19 (2), pp. 50-55.

Panayiotou, A., Kyriakides, L., Creemers, B.P.M., McMahon, L., Vanlaar, G., Pfeifer, M., Rekalidou, G., and Bren, M. (2013). Teacher behavior and student outcomes: Results of a European study. *Paper presented at the American Educational Research Association (AERA) 2013 Conference.* San Francisco, California, April 27- May 1, 2013.

Parducci, A. (1995) *Happiness, pleasure, and judgment: The contextual theory and its applications.* Mahwah, NJ: Erlbaum.

Parr, J.M and Townsend, M.A.R. (2002) Environments, processes and mechanisms in peer learning. *International Journal of Educational Research,* 37, pp.403-423.

Parker, P.D., Marsh, H.W., Lüdtke, O., and Trautwein, U. (in press). Differential school contextual effects for math and english: Integrating the Big-Fish-Little-Pond Effect and the Internal/External Frame of Reference. *Journal of Learning and Instruction*.

Plewis I (1985) *Analysing change: Measurement and explanation using longitudinal data*. New York, Wiley.

Preacher, K. J., Zhang, Z., and Zyphur, M. J. (in press). Alternative methods for assessing mediation in multilevel data: The advantage of multilevel SEM. *Structural Equation Modeling,*18, pp.161-182.

Purkey, W.W. (1970) Self-concept and school achievement. New York: Prentice Hall.

Raudenbush, S.W. (2004) What are value added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics,* 29 (1), pp. 121-129.

Raudenbush, S.W. (1989) The analysis of longitudinal, multilevel data. *International Journal of Educational Research,* 13 (7), pp. 721-740.

Raudenbush, S.W. and Willms, J.D. (1995) The estimation of school effects. *Journal of Educational and Behavioral Statistics*, 20 (4), pp. 307-335.

Ray, A. (2006) *School value added measures in England*. London: DfES.

Raykov, T. and Marcoulides, G.A. (2006) *A First Course Structural Equation Modelling*. (2nd Ed.) Mahwah, NJ: Lawrence Erlbaum Associates.

Reyna, V.F. and Brained, C.J. (2007) The importance of mathematics in health and human judgment: numeracy, risk communication and medical decision making. *Learning and Individual Differences*, 17, pp. 147-159.

Reynolds, D. and Creemers, B. (1990) School effectiveness and school improvement: A mission statement. *School Effectiveness and School Improvement,* 1 (1), pp. 1-3.

Reynolds. D., Chapman, C., Kelly, A., Muijs, D. and Sammons, P. (2012) Educational effectiveness: the development of the discipline, the critiques, the defence, and the present debate. *Effective Education*, 3 (2), pp.109-127.

Robinson, W.S. (1950) Ecological correlations and the behavior of individuals. *American Sociological Review*, 15, (3), pp. 351-357.

Rosenberg, M. (1965). *Society and the adolescent self-image*: Princeton University Press Princeton, NJ.

Rosenberg, M. and Simmons, R. G. (1971) *Black and white self-esteem: The urban school child*. Washington, D.C.:American Sociological Association.

Rossi, Freeman and Lipsey (2004) *Evaluation: A Systematic Approach.* London: Sage.

Rothstein, J. (2009) Student sorting and bias in value added estimation: Selection on observables and unobservables. *Education Finance and Policy,* 4 (4), pp. 537-571.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63, pp. 581–592.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, pp. 473–489.

Rutter, M., Maughan, B., Mortimore, P. Ouston, J. and Smith, A. (1979) *Fifteen Thousand Hours: Secondary Schools and their Effects on Children.* Cambridge, MA: Harvard University Press.

Sammons, P. (1996) Complexities in the judgment of school effectiveness. *Educational Research and Evaluation,* 2 (2), pp. 113-149.

Sammons, P. (1999) *School effectiveness: Coming of age in the twenty-first century.* Lisse, the Netherlands: Swets and Zaitlinger.

Sammons, P. and Bakkum, L. (2011) Effective Schools, Equity and Teacher Effectiveness: A Review to the Literature. Profesorado: Revista de Curriculum y Formaciòn del Profesorado, 15 (3), pp. 10-26.

Sammons, P., Nuttall, D., Cuttance, P. and Thomas, S. (1995) Continuity of school effects: A longitudinal analysis of primary and secondary school effects on GCSE performance. *School Effectiveness and School Improvement,* 6 (4), pp. 285-307.

Särndal, C.E., Swensson, B. and Wretman, J. (2003) *Model assisted survey sampling*. New York: Springer Verlag.

Schafer, J.L. and Graham, J.W. (2002) Missing data: our view of the state of the art. *Psychological methods,* 7 (2), pp. 147-177.

Scheerens, J. and Bosker, R.J. (1997) *The foundations of educational effectiveness.* Oxford: Pergamon Press.

Schochet, P.Z. (2009) Statistical power for regression discontinuity designs in education evaluations. *Journal of Educational and Behavioral Statistics,* 34 (2), pp. 238-266.

Searle, S.R., Casella, G. and McCulloch, C.E. (1992) *Variance components.* New York:Wiley.

Seaton, M. and Marsh, H.W. (2012, July) Celebrating methodological-substantive synergy: Self-concept theory and methodological innovation. Chapter based in part on a keynote presentation at the 2009 SELF Conference held in Quebec, Canada.

Seaton, M., Marsh, H.W. and Craven, R.G. (2010) Big-Fish-Little-Pond-Effect generalizability and moderation-Two sides of the same coin. *American Educational Research Journal,* 47(2), pp. 390-433.

Seaton, M., Marsh, H.W. and Craven, R.G. (2009) Earning its place as a pan-human theory: Universality of the Big-Fish-Little-Pond-Effect (BFLPE) across 41 culturally and economically diverse countries. *Journal of Educational Psychology*, 101, 403-419

Seligman, M.E.P. and Csikszentmihalyi, M. (2000) Positive psychology: an introduction. *American Psychologist,* 55 (1), pp. 5-16.

Shadish, W.R., Cook, T.D. and Campbell, D.T. (2002) *Experimental and quasi- experimental designs for generalized causal inference.* Boston, MA, USA: Houghton Mifflin.

Shavelson, R.J., Hubner, J.J. and Stanton, G.C. (1976) Self-concept: Validation of construct interpretations. *Review of Educational Research,* 46 (3), pp. 407-441.

Shin, Y. and Raudenbush, S.W. (2010) A latent cluster-mean approach to the contextual effects model with missing data. *Journal of Educational and Behavioral Statistics,* 35 (1), pp. 26-53.

Slater, H., Davies, N. and Burgess, S. (2009) *Do teachers matter? Measuring the variation in teacher effectiveness in England.* (Working Paper No. 09/212). Retrieved from University of Bath, The Centre for Market and Public Organization website: http://www.bris.ac.uk/cmpo/publications/papers/2009/wp212.pdf

Smith, M.S. (1972) *Equality of educational opportunity: The basic findings reconsidered.* Center for Educational Policy Research: Harvard Graduate School of Education.

Smith, D., and Tomlinson, S. (1989) *The school effect.* Lancaster: Policy Studies Institute, University of Lancaster.

Snijders, T.A.B. and Bosker, R.J. (2004) *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* London: SAGE.

Steiger, J.H. and Lind, J.C. (1980) Statistically based tests for the number of common factors. Paper session presented at the Annual Meeting of the Psychometric Society, Iowa City, IA.

Stevenson, H.W. and Newman, R.S. (1986) Long-Term Prediction of Achievement and Attitudes in Mathematics and Reading. *Child development,* 57 (3), pp. 646-659.

Strand, S. (1997) Pupil progress during Key Stage 1: a value added analysis of school effects. *British educational research journal,* 23 (4), pp. 471-487.

Summers, A.A. and Wolfe, B.L. (1977) Do schools make a difference? *American Economic Review,* 67 (4), pp. 639-652.

Teddlie, C. (2010). The legacy of the school effectiveness research tradition. In A. Hargreaves, A. Lieberman, M. Fullan and D. Hopkins (Eds.). *The second international handbook of educational change.* Dordrecht: Springer

Teddlie, C. and Reynolds, D. (2000) *The international handbook of school effectiveness research.* New York: Falmer Press.

Teddlie, C., Stringfield, S. and Reynolds, D. (1999) Context issues within school effectiveness research. In C. Teddlie and D. Reynolds. *The international handbook of school effectiveness research* (pp.160-187).

Televantou, I., Marsh, H.W., Kyriakides, L., Nagengast, B., Fletcher, J. ad Malmberg, Lars-Erik (in press) Phantom effects in school composition research: consequences of failure to control biases due to measurement error in traditional multilevel models. *School Effectiveness and School Improvement.*

Thomas, S. and Mortimore, P. (1996) Comparison of value-added models for secondary-school effectiveness. *Research Papers in Education,* 11 (1), pp. 5-33.

Thomas, S. and Mortimore, P. (1994) *Report on value added analysis of 1993 GCSE examination results in Lancashire.* London: Institute of Education.

Thomas, S., Sammons, P., Mortimore, P. and Smees, R. (1997) Stability and consistency in secondary schools' effects on students' GCSE outcomes over three years. *School Effectiveness and School Impovement,* 8 (2), pp.169-197.

Thrupp, M. (1999) *Schools Making a Difference: Let's be realistic! School mix, school effectiveness and the social limits of reform.* UK: Open University Press.

Thrupp, M. and Hursh, D. (2006). The limits of managerialist school reform: The case of target setting in England and the USA. In H. Lauder, P. Brown, J.A. Dillabough, and A. H. Halsey (Eds.), *Education, globalization and social change.* Oxford: Oxford University Press.

Thrupp, M., Lauder, H. and Robinson, T. (2002) School composition and peer effects. *International Journal of Educational Research,* 37 (5), pp. 483-504.

Tizard, B., Burgess, T., Francis, H., Goldstein, H., Young, M., Hewison, J. and Plewis, I. (1980) *Fifteen thousand hours: Secondary schools and their effect on children: A discussion.* London: University of London Institute of Education.

Trautwein, U., Gerlach, E. and Lüdtke, O. (2008) Athletic classmates, physical self-concept and free-time physical activity: A longitudinal study of frame of reference effects. *Journal of Educational Psychology,* 100 (4), pp.988-1001.

Trautwein, U., Lüdtke, O., Marsh, H.W., Köller, O. and Baumert, J. (2006) Tracking, grading, and student motivation: Using group composition and status to predict self-concept and interest in ninth-grade mathematics, *Journal of Educational Psychology,* 98 (4), pp. 788-806.

Trochim, W.M.K. (1984) *Research design for program evaluation: The regression-discontinuity approach.* CA: Sage Beverly Hills

Tymms, P. (2001) A test of the big fish in a little pond hypothesis: An investigation into the feelings of seven-year-old pupils in school. *School Effectiveness and School Improvement,* 12 (2), pp. 161-181.

Tymms, P. (1999) *Baseline assessment and monitoring in primary schools: Achievements,mattitudes and value added indicators.* London: David Fulton.

Tymms, P. and Coe, R. (2003) Celebration of the success of distributed research with schools: The CEM Centre, Durham. *British Educational Research Journal,* 29 (5), pp. 639-667.

Tymms, P., Jones, P., Albone, S. and Henderson, B. (2009) The first seven years at school. *Educational Assessment, Evaluation and Accountability,* 21 (1), pp. 67-80.

Tymms, P., Merrell, C. and Henderson, B. (1997): The first year at school: A quantitative investigation of the attainment and progress of pupils, *Educational Research and Evaluation: An International Journal on Theory and Practice*, 3 (2), pp. 101-118

Tymms, P., Merrell, C., Heron, T., Jones, P., Albone, S. and Henderson, B. (2008) The importance of districts. *School Effectiveness and School Improvement*, 19 (3), pp.261-227.

Valentine, J. C., and DuBois, D. L. (2005). Effects of self-beliefs on academic achievement and vice-versa: Separating the chicken from the egg. In H. Marsh and R. Craven (Eds.), *International advances in self research* (Vol. 2, pp. 53–75). US: Information Age Publishing Inc.

Van de gaer, E., Fraine, B.D., Pustjens, H., Van Damme, J., De Munter, A. and Onghena, P. (2009) School effects on the development of motivation toward learning tasks and the development of academic self-concept in secondary education: a multivariate latent growth curve approach. *School Effectiveness and School Improvement,* 20 (2), pp. 235-253.

Van de Grift, W. (2009) Reliability and validity in measuring the value added of schools. *School Effectiveness and School Improvement,* 20 (2), pp. 269-285.

Vennemann, M. and Wendt, H. (2012). The ADDITION *achievement results: Cross-sectional and longitudinal results of mathematics and science in Cyprus.* Institut für Schulentwicklungsforschung.

Verarchert, P., Van Damme, J., Onghena, P. Ghesquière, P. (2009) A seasonal perspective on school effectiveness: Evidence from a Flemish longitudinal study in kindergarten and first grade. *School Effectiveness and School Improvement*, 20 (2), pp.215-233.

Warner, R.M. (2008) *Applied statistics: From bivariate through multivariate techniques.* London : SAGE.

Wedell, D.H. and Parducci, A. (2000) Social comparison: Lessons from basic research on judgment. In J. Suls and L. Wheeler (Eds.), *Handbook of social comparison: Theory and research.* (pp.223-252). Dodrecht, Netherlands: Kluwer Academic.

Webb, N. M. and Shavelson, R. J. (2005). Generalizability theory: Overview. In B. S. Everitt and D. C. Howell, (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 717-719). London: Wiley.

Whetton, C. (2009) A brief history of a testing time: national curriculum assessment in England 1989-2008. *Educational Research*, 51 (2), pp.137-159.

Widaman, K. F., and Reise, S. P. (1997) Exploring the measurement invariance of Book psychological instruments: Applications in the substance use domain. In K. J. Chapter Bryant, M. Windle, and S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281-324). Washington, DC: American Psychological Association.

Wiley, D.E. and Harnischfeger, A. (1974) Explosion of a myth: Quantity of schooling and exposure to instruction, major educational vehicles. *Educational Researcher*, 3 (4), pp. 7-12.

Wilkinson, I.A.G. (2002) Introduction: peer influences on learning: what are they? *International Journal of Educational Research*, 37, pp.395-401.

Wilkinson, I.A.G., Hattie, J.A., Parr, J.M., Townsend, M.A.R., Fung, I., Ussher, C., Thrupp, M., Lauder, H. and Robinson, T. (2000) *Influence of peer effects on learning outcomes: A review of the literature.* Auckland, New Zealand: Auckland UniServices Limited.

Wilkinson, I.A.G., Parr, J.M., Fung, I.Y.Y., Hattie, J.A.C. and Townsend, M.A.R. (2002) Discussion: modeling and maximizing peer effects in school. *International Journal of Educational Research,* 37 (5), pp. 521-535.

Wilkinson, L. and Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist,* 54 (8), 594-604.

Will, J.D. (1992) *Monitoring school performance: A guide for educators.* London: The Falmer Press.

Williams, R. (2014) *Graduate Statistics II.* Personal Collection of Richard Williams, University of Notre Damme, Notre Damme, Inidana. Retrieved from: http://www3.nd.edu/~rwilliam/xsoc63993/l21.pdf

Willms, J.D. (1985a) Catholic-school effects on academic achievement: New evidence from the High School and Beyond follow-up study. *Sociology of Education,* 58 (2), pp. 98-114.

Willms, J.D. (1985b) The Balance Thesis: contextual effects of ability on pupils' O-grade examination results. *Oxford Review of Education,* 11 (1), pp. 33-41.

Willms. J.D. (1986) Social segregation and its relationship to pupils' examination results in Scotland. *American Sociological Review,* 51 (2), 224-241. Retrieved from: http://www.jstor.org/stable/2095518.

Wilson, A.B. (1959) Residential segregation of social classes and aspirations of high school boys. *American Sociological Review,* 24 (6), pp. 836-845.

Woodhouse, G., Yang, M., Goldstein, H. and Rasbash, J. (1996) Adjusting for measurement error in multilevel analysis. *Journal of the Royal Statistical Society. Series A (Statistics in Society),* 159 (2), pp. 201-212.

Yang, M., Goldstein, H., Rath, T. and Hill, N. (1999) The use of assessment data for school improvement purposes. *Oxford Review of Education,* 25 (4), pp. 469-483.

Xu, M. K. (2010). *Frame of reference effects in academic self-concept: An examination of the Big-Fish-Little-Pond Effect and the Internal/External Frame of Reference model for Hong Kong adolescents*. PhD Thesis, University of Oxford.

Zirkel, P.A.(1971) Self-concept and the "disadvantage" of ethnic group membership and mixture. *Review of Educational Research*, 41, pp. 211-225.