



# Peer Spillover and Big-Fish-Little-Pond Effects with SIMS80: Revisiting a Historical Database Through the Lens of a Modern Methodological Perspective

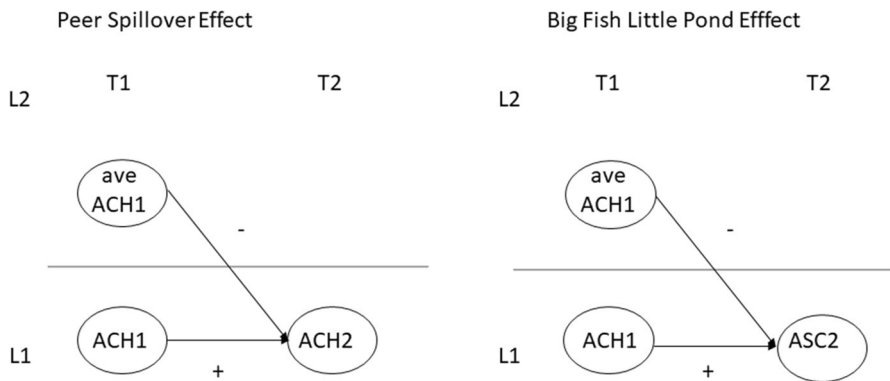
Ioulia Televantou<sup>1,2</sup> · Herbert W. Marsh<sup>3,4</sup> · Kate M. Xu<sup>5</sup> · Jiesi Guo<sup>3</sup> · Theresa Dicke<sup>3</sup>

Accepted: 9 September 2023 / Published online: 5 October 2023  
© The Author(s) 2023

## Abstract

The present study uses doubly latent models to estimate the effect of average mathematics achievement at the class level on students' subsequent mathematics achievement (the “Peer Spillover Effect”) and mathematics self-concept (the “Big-Fish-Little-Pond-Effect; BFLPE”), controlling for individual differences in prior mathematics achievement. Our data, consisting of 13-year-old students from Canada, the USA, and New Zealand, come from a unique cross-national database with a longitudinal design at the student level: the Second International Mathematics Study (SIMS80). This historical survey was administered by IEA in the 1980s and highly influenced the development of educational policies in the following decades. We replicate a widely cited study based on SIMS80, interrogating the validity of its findings of a positive peer spillover effect. When we adjust for measurement error, using doubly latent models, we observe that originally positive peer spillover effects become less positive or disappear altogether. On the contrary, negative BFLPEs become more negative and remain statistically significant throughout. Our study is the only cross-national study to have evaluated both the BFLPE and the peer spillover effect with controls for a true measure of prior achievement — and the only study to test the peer spillover effect cross-nationally using doubly latent models. Our findings question the empirical results of past and current research evaluating school- and class-level compositional effects based on sub-optimal models that fail to control for measurement error.

**Keywords** Measurement error · Doubly latent models · SIMS80 · Compositional effect · Peer spillover effect · Big-fish-little-pond-effect



**Fig. 1** Theoretical model of the peer spillover effect, the Big-Fish-Little-Pond-Effect, and the reciprocal effects model. T1, pre-test (i.e., beginning of the academic year); T2, post-test (i.e., end of the academic year); ACH1, student's academic achievement at T1; ACH2, students' academic achievement at T2; ASC2, academic self-concept at T2; ave ACH1, students' average achievement at T1

## Introduction

Between-class or school differences in the achievement composition of their students may reflect the selective allocation of students in different learning groups according to their academic capabilities (e.g., tracking, ability-grouping, streaming; Ireson et al., 2001), or they may simply be the consequence of social segregation and neighborhood effects, and, even, the result of parental choice. A compositional effect is revealed when students' outcomes are associated with the aggregated characteristics of their peers in the school or the classroom after controlling for pre-existing differences at the student level (Epple & Romano, 2011; Harker & Tymms, 2004; Marsh, Pekrun et al., 2018; 2023; Wagner, 2022). For instance, a positive compositional effect of average achievement would suggest that students of the same academic achievement benefit more if they attend an institution or a classroom with a higher achievement intake. On the contrary, an absent or a negative compositional effect would suggest that attending a higher-achieving institution might not necessarily benefit student learning (Televantou et al., 2015). Research findings often support what is taken to be the conventional wisdom, suggesting a positive, albeit weak effect of class- or school-aggregated achievement on students' academic outcomes (Teddle et al., 1999; Wagner, 2022; Willms, 1985) — the so-called *peer spillover effect* (Cooley Fruehwirth, 2013; see Fig. 1). Thus, they suggest a positive association between the peers' average achievement and a student's academic achievement. Empirical evidence that supports this view can be traced back to the 1960s, to the “Coleman Report” on educational opportunity (Coleman et al., 1966). Since then, many other studies that looked at data from different countries and used different analytical strategies have also shown positive compositional effects on students' academic achievement development (Becker et al., 2021; Burns & Mason, 2002; Hanushek et al., 2003; Nomi & Raudenbush, 2016; Opdenakker et al., 2002). At the same time,

another set of studies reports negative or non-existent compositional effects (for an overview, see Sacerdote, 2011). In fact, there is remarkably little agreement on this matter (Ammermueller & Pischke, 2009; Hanushek et al., 2003; Sacerdote, 2001; Stinebrickner & Stinebrickner, 2006). Variation in the reported findings is linked to, among other factors, the inadequacies in different methodologies used and the quality of the data collected (Hutchison, 2004; Manski, 1993; Thrupp et al., 2002). Recent empirical (Harker & Tymms, 2004; Hutchison, 2007; Marsh et al., 2010; Televantou et al., 2015; Woodhouse et al., 1996) and methodological work (Marsh et al., 2009; Pokropek, 2015) has specifically turned its focus on the impact of correcting for measurement error in student-level measures (i.e., student achievement) on compositional effects estimates. Based on a longitudinal sample of US children who participated in the Early Childhood Longitudinal Study Kindergarten Class of 1998-99, Dicke et al. (2018) demonstrated that when the appropriate methodology is used, which allows for adjustments for measurement error in achievement scores, and controls for pre-existing differences, originally positive school-level compositional effects become minimal, close to zero. The authors called for further studies to investigate their findings' replicability (Nosek et al., 2022).

Dicke and colleagues addressed the importance of evaluating compositional effects concerning academic achievement on another educational outcome, namely, academic self-concept (see also Stäbler et al., 2017; Televantou et al., 2021). Considering the positive effect of average achievement on students' academic achievement and the reciprocal effects model, which suggests a mutually positive relationship between student achievement and student self-concept, one would expect a positive effect of average achievement on students' academic self-concept. However, consistent evidence in the educational psychology literature suggests a negative relationship between average achievement and students' self-concept (Marsh & Craven, 2005; Marsh & Martin, 2011), the *Big-Fish-Little-Pond-Effect* (BFLPE; Marsh, 1987; Marsh, Xu et al., 2021; see Fig. 1). Dicke et al. claimed, and empirically showed, that correcting for measurement error and pre-existing student differences may be the key to achieving convergence between BFLPE research findings of negative compositional effects on self-concept and the educational research findings of positive compositional effects on achievement. They based their suggestion on their finding that negative school-level BFLPEs turned even more negative after such adjustments. In contrast, positive school-level compositional effects became much smaller, slightly below zero. In this respect, the inconsistency in the estimates of the two compositional effects was eliminated, and the aforementioned theoretical paradox was partially resolved. In a subsequent study, however, Becker et al. (2021) suggested that the mixed pattern in studies investigating compositional effects on achievement and self-concept in educational settings might not be an artifact of inadequate methodology alone; it might instead reflect a "substantive effect pattern" (Becker et al., 2021; p. 14). They propose that researchers should look into the mechanisms driving the occurrence of compositional effects—in addition to using the appropriate research methodology—before they conclude the appropriateness of statistical findings. Becker et al. distinguished between different mechanisms

leading to the occurrence of achievement composition effects, namely peer processes and instructional processes, as well as the allocation of resources to schools or classrooms. They explain that factors associated with these factors should be controlled for in statistical analysis before inferences about the actual size of achievement composition effects are made.

The present investigation revisits historical data from the 1980s, administered by the International Association for the Evaluation of Educational Achievement (IEA), namely the Second International Mathematics Study (SIMS80). Based on these data, we reproduce a positive and statistically significant class-level compositional effect of mathematics achievement (Zimmer & Toma, 2000). However, this effect largely disappears when controls for measurement error are made, in line with the recommendations of more recent studies (Dicke et al., 2018; Televantou et al., 2015, 2021), and controlling for a range of class-level variables, in line with Becker et al. (2021). Thus, the original study of Zimmer and Toma, despite being highly cited (Hanushek et al., 2003; Hanushek & Woessmann, 2011; Sacerdote, 2011; Van de Werfhorst & Mijs, 2010), fails to be replicated (Nosek et al., 2022) with a more appropriate analytical approach. Further, the present study evaluates the compositional effect of mathematics achievement on mathematics self-concept (Dicke et al., 2018; Stäbler et al., 2017), demonstrating the robustness of the BFLPE for different analytical strategies. The value of doing and understanding replication, reproducibility, and robustness has been increasingly recognized in the past decade as contributing to the quality of research findings and accelerating scientific progress (Nosek et al., 2022); our study involves aspects of all these three elements.

The Second International Mathematics Study (SIMS80) and FIMS, the First International Mathematics Study administered in the 1960s, represent the first two international comparisons of mathematics achievement. While both surveys have substantially influenced education worldwide, SIMS80 provides a more valid basis for relevant empirical studies (Brown, 1996). Significantly, SIMS80 is based on a pre- and post-measurement design, allowing for controls for the effects of prior achievement on subsequent achievement (the peer spillover effect; see above) and the BFLPE. Although this might also be the case for some other studies that involve longitudinal data, ours might be the only study to have this cross-nationally. In particular, all subsequent IEA and PISA data collections have been strictly cross-sectional. Whereas the BFLPE can be tested with cross-sectional data, tests of the peer spillover effect cannot (Caro et al., 2017; Wagner, 2022). In this respect, our study is the only cross-national study to have evaluated both the BFLPE and the peer spillover effect with controls for a true measure of prior achievement — and the only study to test the peer spillover effect cross-nationally using doubly latent models.

The following sections describe how current state-of-the-art compositional analysis models build on Zimmer and Toma's (2000) analytical approach, adjusting for measurement error bias in compositional effect estimates. We then justify why we considered academic self-concept as an educational outcome, in addition to mathematics self-concept, describing the Big-Fish-Little-Pond-Effect (BFLPE). Finally, we give our study's scope, research hypotheses, and research questions.

## State-of-the-Art Compositional Analysis Models

In estimating the class-level compositional effect of average achievement, Zimmer and Toma (2000) used a fixed effects regression model linear-in-means model (Ammermueller & Pischke, 2009); this is a typical approach being followed in the econometrics literature. Today's default approach to compositional analysis in education is multilevel modeling (Snijders & Bosker, 2012). With both multiple regression linear-in-means models and multilevel compositional analysis models, the criterion, typically student-level performance in an outcome of interest (academic self-concept, academic achievement), is regressed on an individual-level variable (prior achievement), and the corresponding class- or school-level aggregate (average achievement). The effect of the aggregated variable on the student-level outcome is commonly referred to as the *compositional* effect (Harker & Tymms, 2004). Multilevel modeling's strength lies in the fact that it accounts for the nesting in the structure of educational data (e.g., students nested into classrooms or schools), providing unbiased standard errors — typically, with multiple regression models, standard errors are estimated larger. However, they involve single-scale scores (manifest concerning the sampling of items) and manifest aggregation (manifest concerning the sampling of people). Marsh et al. (Marsh et al., 2009; Marsh & Martin, 2011; Marsh et al., 2012) developed and demonstrated the application of *doubly latent* models (latent variable, latent aggregation) to investigate compositional effects. This approach builds on the multilevel model, referred to as the *doubly manifest* model (Marsh et al., 2009), by allowing controls for measurement error in individual-level measures and corresponding aggregates and sampling error in the aggregated measures. In this methodological framework, measurement error is conceptualized as the result of using only a finite number of items to measure a student's academic achievement or trait. In contrast, an infinite number of items would have been needed to obtain a reliable measurement. It is controlled by using multiple indicators (Marsh et al., 2009). Sampling error arises when only a finite number of individuals from each higher-level unit is used to form the aggregated measures. It is adjusted for by using latent rather than a manifest aggregation to form the group-level aggregates (Lüdtke et al., 2008).

## Phantom Peer Spillover Effects

Harker and Tymms (2004) coined the term *phantom effects* to describe misleadingly positive effects of aggregated variables in compositional models — peer spillover effects that are simply an artifact of the inadequacy in the statistical procedures used. Two “facets” (Televantou et al., 2015, p. 79) of under-specification at the student level have been shown to lead to the so-called phantom effects: measurement error bias and omitted variable, or selection bias (Caro et al., 2017; Harker & Tymms, 2004; Televantou et al., 2015). Omitted variable bias refers to insufficient controls for student-level background measures — a problem common to all observational studies (Pearl, 2002; West & Thoemmes, 2010). Measurement error bias may lead to positive compositional effects misleadingly appearing as more positive and non-existent ones being estimated as positive and significant. Conversely, negative compositional effects may be estimated

as less negative, and in the presence of large amounts of measurement error, they may even turn positive (Televantou et al., 2015). The direction of omitted variable bias in the compositional effect estimate is not straightforward to predict: it depends on the correlation between the omitted and the aggregated variable, as well as on the relative direction of the effects of the two variables on the outcome of interest (Caro et al., 2017). Previous research (Dicke et al., 2018) has empirically shown that correcting for omitted variables at the student level eliminates the peer spillover effect and leads to a more negative BFLPE. However, the researchers call for further studies to validate their findings with different data and background variables. In their study, Zimmer and Toma (2000) dealt with omitted variable bias through controls for a diverse range of student-level characteristics available with SIMS80. However, they could not deal with measurement error bias reasonably since statistical models that can accommodate this source of bias were not readily available at the time their study took place.

### The Big-Fish-Little-Pond-Effect

Academic Self-Concept (ASC) is defined as the specific component of self-concept that denotes how individuals perceive their academic abilities and competencies in a specific subject (Byrne & Shavelson, 1986). ASC is valued as an educational outcome in its own right (Zirkel, 1971) and as a facilitator of other desirable outcomes (Guay et al., 2004; Ivanova & Michaelides, 2022). Significantly, ASC and academic achievement are reciprocally related — the reciprocal effects model (REM; Guo et al., 2018; Marsh, 2023; Marsh & Craven, 2005, 2006; Marsh & Martin, 2011; Marsh, Pekrun et al., 2022 — so that higher ASC facilitates higher academic achievement and vice versa. BFLPE studies center on the consequences of attending a high-achieving classroom or school on academic self-concept (Fig. 1). They show that students with similar academic achievement levels feel less competent in high-achieving classrooms than in average- or low-achieving ones (Marsh, 1987). The theoretical explanation for the BFLPE is based on social comparison theory (Festinger, 1954; Marsh, Xu et al., 2021), which emphasizes the need to consider the relative frames of reference to understand how people perceive their competencies in specific domains (Marsh et al., 2014; Marsh, Pekrun et al., 2018). The negative BFLPE is conceptualized as the net effect of two processes (Marsh et al., 2000): a positive assimilation effect due to being affiliated with a prestigious institution or a highly selective educational program, and a negative contrast due to social comparisons with higher-achieving peers.

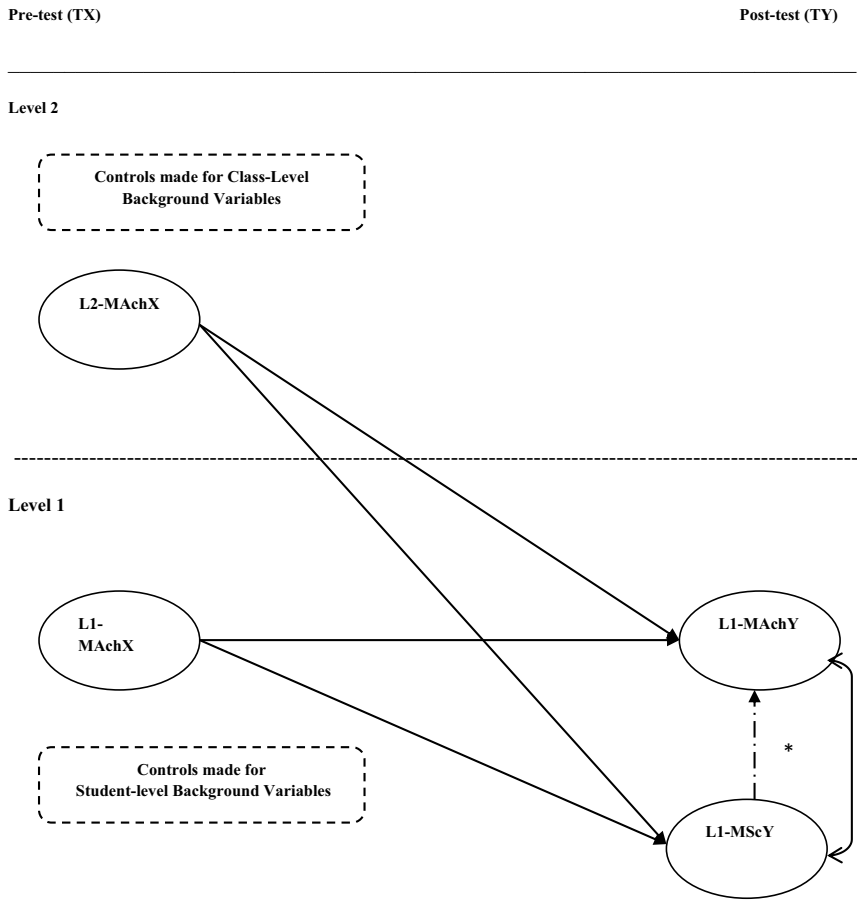
The BFLPE is one of psychology's most cross-culturally universal phenomena, verified with three successive PISA data collections (Marsh & Hau, 2003; Marsh, Xu et al., 2021; Seaton et al., 2009). It generalizes across student groups: subject domains, ASC instruments, and cultures (Basarkod, et al., 2023; Marsh et al., 2008; Marsh et al., 2015; Nagengast & Marsh, 2012), and has been shown using standardized test scores and school grades (Marsh, Pekrun et al. 2022; also see Fleischmann et al., 2021), and in terms of class rank instead of class-average achievement (Loyalka et al., 2018; Marsh et al., 2020). The present study aims to verify the hypothesis that BFLPEs remain negative and statistically significant after measurement error is controlled for (Dicke et al., 2018; Televantou et al., 2021).

## Class-level Confounders of the “Pure” Compositional Effect

Investigating compositional effects in educational settings is closely linked to whether “Schools Matter” (Mortimore et al., 1988). If the effects are substantial, then researchers interpret this as an indication that students’ achievements are influenced by the interaction of students with each other in the school’s social context (Thrupp et al., 2002). In assigning such interpretations to compositional effect estimates, however, empirical researchers must be cautious; compositional effects may reflect something more than simply “peer contagion” (Dishion & Tipsord, 2011) — the influence that students may exert on each other. Differences in average achievement across classrooms or educational institutions are typically confounded with inequalities in instructional processes and allocation of resources (Thrupp, 1999). For example, teachers respond to the group of students they teach, and vice versa; high-achieving classes may attract teachers with more teaching experience and a higher level of pedagogical training (Fauth et al., 2021). Concerning these issues, Becker et al. (2021) distinguished between “general” and “pure” compositional effects, the latter being a more accurate approximation of the actual peer effect. “Pure” compositional effects, or peer effects, can be approximated by adjusting for the effects of extra-compositional variables that act as potential confounders of the effect of aggregated characteristics at the class or the school level (Dicke et al., 2018; Marsh, Pekrun et al., 2023; Televantou et al., 2015). In our study, following Zimmer and Toma (2000), we exploit the richness of SIMS80 data, controlling for a range of class-level extra-compositional variables (e.g., the teacher’s experience and pedagogical knowledge) in compositional analysis models. This way, we aim to approximate the “pure” compositional effect better.

## The Present Study

The present study initially attempts to reproduce the peer spillover effect (Fig. 1) reported by Zimmer and Toma (2000). using a subset of the data used in their analysis (see the “Data Sample and Measures” section and Supplementary Materials). Our research hypothesis (Research Hypothesis 1/RH1) is that a positive class-level compositional effect of average achievement on subsequent mathematics achievement will be retrieved, in line with the original study. Second, we replicate the study of Zimmer and Toma, using doubly latent models that correct for measurement error. We anticipate that peer spillover effects will become less positive or disappear once measurement error bias is adjusted (Research Hypothesis 2/RH2; Lüdtke et al., 2008; Marsh et al., 2009; Televantou et al., 2015). Additionally, we test the peer spillover effect and the BFLE simultaneously, integrating the two effects in one path analyses model (Fig. 2; Dicke et al., 2018; Televantou et al., 2021; Stäbler et al., 2017). Following our analysis of the integrated model, we observe the amount of bias in the compositional effect estimates, the peer spillover effect, and the BFLPE, due to different forms of model misspecifications — not controlling for measurement error, omission of student-level variables, and omission of class-level variables



**Fig. 2** Conceptual model of the paths tested in the present study. L1-MAchX corresponds to individual mathematics achievement at pre-test, L2-MAchX to mathematics achievement aggregated at level 2, L1-MAchY to mathematics achievement at post-test, and L1-MScY to mathematics self-concept at post-test. Level 2 is the class. Dashed lines represent controls for student background characteristics and additional controls for class-level extra-compositional effects. The range of background variables selected at both levels was based on Zimmer and Toma (2000). \*Path tested in an exploratory study assuming mediation of the effect of the compositional effect of class-average achievement on subsequent achievement, through mathematics self-concept

— one in isolation from the other. We expect the negative and statistically significant BFLPE to be robust to different model specifications (Research Hypothesis 3/RH3), in line with Dicke et al. (Dicke et al., 2018; Televantou et al., 2021). Peer spillover effects are expected to be estimated more positive when measurement error is not adjusted for (Becker et al., 2021; Dicke et al., 2018; Harker & Tymms, 2004; Pokropek, 2015; Televantou et al., 2015). Moreover, they are expected to be sensitive to the set of background variables controlled for in the models (Research Hypotheses 4/RH4). Finally, in an analysis of exploratory nature, we assume mediation



of the effect of class-level effect of average mathematics achievement on students' subsequent mathematics achievement through self-concept (Fig. 2; see dashed line). We focus on the significance and direction of the mediating effect of mathematics self-concept in the relationship between class-average achievement and subsequent mathematics achievement. Our rationale is that if a negative and statistically significant mediation is revealed, this would indicate that the BFLPE and the peer spillover effects might not operate independently.

## Methodology

### Data Sample and Measures

Descriptive measures of our data, a large sample of 13-year-old students from the USA, Canada (Ontario), and New Zealand, are given in Table 1 (see "A.1 Data Samples" in Supplementary Materials).

### Mathematics Achievement

Following Zimmer and Toma (2000), we based our mathematics achievement measures on forty items present in all four distinct rotated forms of the SIMS80 mathematics tests (Zimmer & Toma, 2000). The items were common in pre- and post-measurement occasions; each item was given a value of 1 if it was answered correctly and a value of 0 if it was answered wrongly, and was left blank if no answer was given in the specific question. They were a mixture of arithmetic and word problems covering a range of mathematical topics (algebra, geometry). We first used multiple imputations to treat missingness at the item-level (see "Missing Data" section) and then used item parceling (Little et al., 2002, 2013, 2022) to form the multiple indicators for the pre- and post-test. For both measures, we created four 10-item parcels, taking the average of every 4th item available, allowing for more parsimonious statistical models to be used (Marsh et al., 2013). Importantly, concerning the purposes of our analysis, in which the original items were dichotomous, item parcels gave indicators with a distribution closer to normal, thereby facilitating normal theory-based estimation (Marsh et al., 2013). Manifest measures of achievement were based simply on each student's average score. All variables at level 1 were standardized by subtracting the overall mean and dividing by the overall standard deviation to have a mean of 0 and a standard deviation of 1.

### Mathematics Self-concept

Mathematics self-concept measures were only present at the end-of-year exams. However, four items were used as indicators of students' mathematics self-concept, namely, "I could never be a good mathematician," "I am not so good at Mathematics," and "I cannot do well at Mathematics, no matter how hard I try" and "Mathematics is harder for me than for most people." The items were on a 5-point agreement Likert scale (1 = "Strongly Disagree" ... 5 = "Strongly Agree"), but they

**Table 1** Descriptive analysis for pre- and post-test measures of mathematics achievement and mathematics self-concept for the three countries in our sample, USA, Canada (Ontario), and New Zealand

Country	Number of students	Number of schools	Average school size	Number of classes	Average class size	Intra-class correlation coefficient						Reliability McDonald's omega ( $\omega$ )		
						Math Pre-test	Math Post-test	Math Self-concept	Math Pre-test	Math Post-test	Math Self-concept	Math Pre-test	Math Post-test	Math Self-concept
USA	4857	129	38	225	21.58	.572 <sup>a</sup>	.559 <sup>a</sup>	.105 <sup>a</sup>	.884	.912	.831	.884	.912	.831
Canada (Ontario)	4100	111	37	175	23.43	.265 <sup>a</sup>	.265 <sup>a</sup>	.027 <sup>a</sup>	.853	.886	.838	.853	.886	.838
New Zealand	4699	100	37	199	23.61	.499 <sup>a</sup>	.482 <sup>a</sup>	.080 <sup>a</sup>	.881	.902	.856	.881	.902	.856
Total	13,656	340												

<sup>a</sup>Level 2 is the classroom

were reversely coded, so that scores closest to 5 would reflect higher mathematics self-concept.

### Reliability Estimates

Measurement error reliability was relatively high for mathematics achievement and self-concept measures — McDonald's omega ( $\omega$ ) was estimated higher than .8 (see Table 1). Thus, measurement error in our data was not substantial.

### Student Background Measures

Our models included student-level variables representing family and socio-economic characteristics (Table A.1, Supplementary Materials). We considered the same sets of student background variables as the reference study (Zimmer & Toma, 2000): father's and mother's occupation (un- or semiskilled, skilled worker, clerical or sales, professional), father's and mother's higher education attained (little or none, primary school, secondary school, beyond secondary), students' gender (female, male), and the frequency with which the language of instruction is spoken at home (never, sometimes, usually, always).

### Class-Level Background Measures

Class-level variables considered in our analysis were mainly relevant to the teacher's characteristics (Table A.1, Supplementary Materials). More specifically, we considered the teacher's gender (male/female), the number of years of teaching experience, the number of years teaching mathematics to 8-year students, the number of courses in mathematics methods and pedagogy that were included in the teacher's post-secondary education, and the number of courses in general methods and pedagogy that were included in the teacher's post-secondary education. Finally, we controlled for the type of school (Private or Public) in which the class was situated, the kind of community served by the school (Rural, Suburban, Urban/Suburban, Inner-city metropolis<sup>1</sup>), and the total number of students enrolled in the school.

## Statistical Analysis

### Missing Data

The percentage of missing data for the different variables involved in our analysis was not substantial (Table A.1; Supplementary materials). We used multiple imputations to treat missing data (Rubin, 1987; Schafer & Graham, 2002); imputation procedures were run in IBM SPSS Statistics (IBM SPSS Missing Values 21, 2012). The procedure involved replacing missing values with a list of five simulated

---

<sup>1</sup> For cities with a population greater than half a million.

values. Missing items in the mathematics self-concept scale were imputed in the same imputation model as mathematics achievement items.. The procedure involved replacing missing values with a list of five simulated values. Each plausible version of the complete data was analyzed using a complete data method. The results were combined to obtain overall estimates and standard errors in the statistical package Mplus 8 (Muthén & Muthén, 2017). Where relevant, class-level variables were computed based on the imputed data.

## Statistical Models

We first replicated the original analysis by Zimmer and Toma, using a multiple regression linear-in-means model and specifying interaction terms of class-average achievement with the dummy variables for the three countries considered. In models controlling for measurement error, we used a multilevel latent variable analysis framework, and we specified multi-group doubly latent models with the grouping variable being the country. The models were applied in Mplus 8 (Muthén & Muthén, 2017). Before any analyses, we established the invariance of the factor structure of the latent factors (mathematics achievement, mathematics self-concept; Raykov et al., 2013) across countries to facilitate meaningful interpretation of observed differences (see Tables A.3; A.4 in Supplementary Materials).

## Effect Size Measures

To facilitate comparisons of the effects estimated across different modeling approaches, and with previous research findings, effect sizes (*ESs*) were calculated according to the recommendations of Marsh et al. (2009) using the following formula:

$$ES_{\beta_{com}} = 2 * \beta_{com} * SD_{com} / \sigma_e \quad (1)$$

We used Eq. (1) to calculate the effect size of the effect of the aggregated variable: what we refer to as the compositional effect ( $ES_{\beta_{com}}$ ). The denominator for both equations is the same, the level 1 residual standard deviation of the score of the students in year 4 mathematics score — or of the score for students' academic self-concept in the case of the BFLPE. The unstandardized regression coefficient,  $\beta_{com}$ , is multiplied by the standard deviation of the predictor. This effect size is interpreted as the difference in the dependent variable between two classes that differ by two standard deviations on the predictor variable.

## Model Fit

For assessing the fit of our models, we used sample size independent fit indices (Marsh et al., 2015): the Tucker–Lewis index (TLI), and the Comparative Fit Index (CFI) — these vary along a 0–1 continuum, and values greater than .90 and .95 typically reflect acceptable and excellent fit to the data, respectively. We also used the

**Table 2** Replication of Zimmer and Toma (2000) study: Using a multiple regression model vs a multilevel latent variable model, and, SIMS80 data from the USA, Canada (Ontario), and New Zealand to estimate the class-level compositional effect of average mathematics achievement on students' mathematics achievement at the fourth year

Multiple regression model	Estimate (standard error)
Average pre-measure	.235 (.027)*** .158 (.018)***
Average pre-measure * country	-.008 (.011) -.012 (.017)
Doubly latent model	Estimate (standard error)
USA	.040 (.036) .164 (.147)
Canada (Ontario)	.119 (.069) .229 (.125)
New Zealand	-.035 (.029) -.210 (.181)

$p < .05$ , \*\*\* $p < .001$

*Standardized effects* are reported in *italics*. For multiple regression, they are based on the STDYX standardization in Mplus. For doubly latent models, the effect size is reported. The effects of background variables are reported in Table A.2. in Supplementary Materials

Root-Mean-Square Error of Approximation (RMSEA), with values of less than .05 and .08 reflecting a close fit and a minimally acceptable fit to the data, respectively.

## Results

### Failing to Replicate Past Findings on the Existence of Peer Spillover Effects

As expected (RH1) and consistent with the original analysis by Zimmer and Toma (2000), applying a multiple regression linear-in-means model led to a positive and statistically significant compositional effect of class-average achievement. In addition, a non-statistically significant interaction term between the country with class-average achievement (see Table 2) was retrieved, suggesting a positive compositional effect for all three countries involved with our data, the USA, Canada (Ontario), and New Zealand. However, when we replicated the study of Zimmer and Toma, using a multilevel latent variable framework, we failed to retrieve the positive peer spillover effect — in line with RH2. More specifically (Table 2), the positive class-level compositional effect of average mathematics achievement detected in study 1 was eliminated. It became non-statistically significant for the two countries involved (USA, New Zealand).

### Modeling the BFLPE and the Peer Spillover Effect in an Integrated Model

The theoretical path models underlying the formation of the BFLPE involve a pre-test in individual achievement and the corresponding school- or class-average achievement as a covariate at level 1 and level 2, respectively (Dicke et al., 2018; Stäbler et al., 2017). Mathematics self-concept at post-test is the outcome. In our analyses, we tested the peer spillover effect and the BFLE simultaneously,

**Table 3** Quantifying bias in the peer spillover effect and the BFLPE due to different forms of model misspecification

Country	Model	Big-Fish-Little-Pond-Effect <sup>1</sup>	Peer spillover effect <sup>1</sup>
USA	doubly manifest	-.294 (.063)***	.808 (.095)***
	doubly latent	-.657 (.091)***	.295 (.144)*
	doubly latent with covariates <sup>2</sup>	-.630 (.091)***	.231 (.143)
	doubly latent with covariates <sup>3</sup>	-.606 (.091)***	.164 (.150)
Canada (Ontario)	doubly manifest	-.424 (.051)***	.376 (.088)***
	doubly latent	-.646 (.071)***	.275 (.128)*
	doubly latent with covariates <sup>2</sup>	-.680 (.078)***	.246 (.132)
	doubly latent with covariates <sup>3</sup>	-.677 (.078)***	.300 (.120)*
New Zealand	doubly manifest	-.453 (.058)***	.646 (.080)***
	doubly latent	-.890 (.091)***	-.023 (.178)
	doubly latent with covariates <sup>2</sup>	-.889 (.096)***	-.116 (.178)
	doubly latent with covariates <sup>3</sup>	-.878 (.097)***	-.238 (.182)

\* $p < .05$ , \*\*\* $p < .001$

<sup>1</sup>Effect size of the corresponding estimate is reported

<sup>2</sup>Adjustments were made for student-level background variables based on Zimmer and Toma (2000): father's and mother's occupation (un- or semiskilled, skilled worker, clerical or sale, professional), father's and mother's higher education attained (little or none, primary school, secondary school, beyond secondary), students' gender (female, male), the frequency with which the language of instruction is spoken at home (never, sometimes, usually, always).

<sup>3</sup>Additional adjustments made for class-level background variables based on Zimmer and Toma (2000): the teacher's gender, number of years of teaching experience, number of years teaching mathematics to 8-year students, number of courses in mathematics methods and pedagogy included in the teacher's post-secondary education, number of courses in general methods and pedagogy that were included in the teacher's post-secondary education, number of students in the target class, number of students being enrolled in the school, type of school (Public, Private), community served by the school (Rural, Suburban, Urban/Suburban, Urban, Inner-city metropolis)

integrating the two models in one path analysis model (Fig. 2). The estimates for the two effects for a fully specified doubly latent model, controlling for measurement error and all background variables, are given in Table 3 (Model: “doubly latent with covariates<sup>3</sup>”). The estimates for the peer spillover effect are essentially the same as those in Table 2, i.e., in the analysis where mathematics achievement is used as the only outcome in the model. The BFLPE was found negative and statistically significant for all three countries — USA ( $\beta_{comp} = -.281$ ,  $SD = .091$ ,  $ES = -.606$ ), Canada (Ontario) ( $\beta_{comp} = -.496$ ,  $SD = .053$ ,  $ES = -.677$ ), and New Zealand ( $\beta_{comp} = -.420$ ,  $SD = .037$ ,  $ES = -.878$ ). Importantly, when constraints of equality for the estimates of the estimated compositional effects were imposed (restricted multi-group doubly latent model; equal compositional effects across countries), we found an overall negative and non-significant compositional effect of class-average achievement on subsequent mathematics achievement ( $\beta_{comp} = -.041$ ,  $SD = .023$ ,  $ES = -.152$ ). The BFLPE detected with the restricted multi-group doubly latent model (equal compositional effects across countries)

was  $\beta_{comp} = -.336$  ( $SD = .023$ ,  $ES = -.693$ ). The Model Fit was not substantially affected by restricting the compositional effects to be equal across countries (Unrestricted Model:  $\chi^2 = 3242.622$ ,  $d.f. = 846$ ,  $RMSEA = .025$ ,  $CFI = .970$ ,  $TLI = .965$ ; Restricted Model:  $\chi^2 = 3879.612$ ,  $d.f. = 866$ ,  $RMSEA = .028$ ,  $CFI = .962$ ,  $TLI = .957$ ).

### Impact of Model Misspecification on Peer Spillover Effect and BFLPE Estimates

Another aim of our study was to quantify the amount of bias in the compositional effect estimates that could be attributable to different forms of model misspecification (failure to control for measurement error, not controlling for appropriate student- and class-level background variables). Relevant findings are displayed in Table 3: The BFLPE remained negative and statistically significant despite the potential bias in statistical estimates, in line with RH3. In contrast, the peer spillover effect estimate was highly vulnerable to the different model specifications, consistent with RH4.

More specifically, *failing to control for measurement error* led to artefactual peer spillover effects. Controlling for measurement error alone only partially corrected for the positive bias in the peer spillover effect; the effect was revealed with Canada (Ontario) and USA data. Peer spillover effects disappeared altogether when *additional controls for student background measures were made*.

Both facets of under-specification at the student level (see the “[Phantom Peer Spillover Effects](#)” section) led to BFLPEs being estimated smaller in magnitude (i.e., less negative); the effects remained statistically significant throughout.

*Failing to include class-level background measures* in our models did not substantially affect our conclusions regarding the magnitude and direction of peer spillover effects and BFLPEs. However, the peer spillover effect was revealed for Canada (Ontario) once controls for class-level background variables were made in the analyses.

### Modeling Mathematics Self-concept as a Mediator

In a supplementary study of exploratory nature, we modeled the mediation of the class-level compositional effect on subsequent mathematics achievement through mathematics self-concept. We found a small, statistically significant negative mediation effect for all three countries (Table 4). When equality restrictions of compositional and mediation effects were imposed for the three countries, the mediation effect was estimated as negative and statistically significant ( $\beta_{med} = -.052$ ,  $SD = .005$ ,  $p < .001$ ). Class-average achievement had a negative effect on mathematics self-concept (BFLPE; Table 4). The results suggest that at least part of the effect of class-average achievement is mediated via mathematics self-concept and this mediated effect is negative.

**Table 4** Estimates of the direct, indirect, and total peer spillover effect, and for the BFLPE

	Peer spillover effect (standard error)	Big-Fish-Little-Pond Effect (standard error)
USA		
Direct effect	.114 (.038)*	
Indirect effect	-.038 (.006)***	
Total effect	.076 (.038) <i>ES</i> = .327	-.281 (.041)*** <i>ES</i> = -.606
Canada (Ontario)		
Direct effect	.284 (.067)***	
Indirect effect	-.120 (.017)***	
Total effect	.164 (.067)* <i>ES</i> = .335	-.496 (.053)*** <i>ES</i> = -.677
New Zealand		
Direct effect	.015 (.032) <i>ES</i> = -.092	
Indirect effect	-.033 (.007)***	
Total effect	018 (.030) <i>ES</i> = -.113	-.420 (.037)*** <i>ES</i> = -.878

\* $p < .05$ , \*\* $p < .05$ , \*\*\* $p < .001$ . *ES* denotes the Effect Size Estimate for the Total Effect. The Direct Effect corresponds to the path from class-level average achievement at pre-test to mathematics achievement at post-test (see Fig. 2). The Indirect Effect corresponds to the path from class-average achievement to mathematics achievement at post-test via mathematics self-concept at post-test (see Fig. 2, dashed line). The Total Effect is the sum of the Direct Effect and the Indirect Effect

## Discussion

The impact of school- or class-average achievement on students' outcomes has received a growing concern among researchers (Becker et al., 2021; Yang Hansen et al., 2022). Despite years of accumulated research, considerable confusion remains about how past analyses can be interpreted. Many relationships reported between student achievement and their class or school peers characteristics are limited by conceptual and analytical shortcomings (Dicke et al., 2018; Harker & Tymms, 2004; Thrupp et al., 2002).

Our study, a large longitudinal sample of students in eighth grade from the USA, Canada, and New Zealand participating in SIMS80, represents another yet scholarly attempt to estimate the class-level compositional effects of achievement on students' mathematics achievement (the peer spillover effect) and students' mathematics self-concept (the Big-Fish-Little-Pond-Effect; BFLPE) using multilevel doubly latent models (Becker et al., 2021; Dicke et al., 2018; Televantou et al., 2021). Zimmer and Toma (2000) used multiple regression linear-in-means models, the state-of-the-art approach at the time their study was conducted. They reported a positive effect of class-average achievement that was statistically significant and robust across the countries in their sample; their findings are not replicated when we apply doubly latent models.



Our findings should not be the sole basis for informing current issues in educational policy and practice as the data are dated, and the generalizability over different countries must be considered. However, they question the empirical results of past and current research evaluating compositional effects based on sub-optimal models that fail to control for measurement error. Importantly, they demonstrate the robustness of the BFLPE to different modeling specifications and datasets used.

The present investigation represents the first cross-national study that simultaneously investigates the BFLPE and the peer spillover effect, controlling for true measures of prior achievement. Students' prior achievement has been shown to explain up to 50% of their differences in subsequent educational outcomes (Colom & Flores-Mendoza, 2007); failing to make such adjustments may lead to overestimating the peer spillover effect (Harker & Tymms, 2004; Wagner, 2022). Further, it evaluates compositional effects at the classroom level and uses data on students in their eighth year of educational studies. This differs from existing studies' focus (Dicke et al., 2018; Televantou et al., 2021), which used doubly latent models to evaluate compositional effects on students' mathematics achievement and mathematics self-concept; their interest was at the level of the school, and they used data from younger students. In general, compositional effects in the educational context are more prominent when looking at the composition of a class rather than that of a school, as the class is the immediate learning environment to which students belong (Marsh et al., 2014). Thus, it is vital to show that positive peer spillover effects can also be artefactual — even at the classroom level — where they would have been expected to be larger, i.e., more positive.

We distinguish pure compositional effects associated with the achievement levels of peers from class-level variables associated with class-level variables that are likely to be confounded with class-average achievement (e.g., the teachers' qualifications). The basic concept (Becker et al., 2021; Hanushek et al., 2003) is that classroom-level fixed effects remove selection effects and allow the researcher to identify peer effects from idiosyncratic variation in peer ability. Becker et al. (2021) made similar arguments and controlled for the effect of tracking in models estimating class-level compositional effects of achievement. In our study, we adjust for a broader range of class-level characteristics. No substantial differences were observed in the compositional effect estimates after such adjustments. Further consideration needs to be done to interpret this finding, considering the educational systems of the countries involved in our sample when the data were gathered. However, this is beyond the scope of our study.

### **Demonstrating Robustness of the BFLPE**

We demonstrate BFLPEs with mathematics self-concept data collected over 40 years ago in the early 1980s. BFLPEs are evident for all three educational systems involved, the USA, Canada (Ontario), and New Zealand. Specifically, with Canada (Ontario), we observe a negative effect of class-average achievement on students' self-concept, despite the apparent weak positive effect of the same compositional variable on students' subsequent mathematics achievement. Adjusting for measurement error and omitted variable bias leads to an even more negative BFLPE, consistent with findings reported by Dicke et al. (2018) and Televantou et al. (2021).

It is essential to understand why controlling for measurement error and covariates strengthens the BFLPE but weakens or eliminates the peer spillover effect. The explanation is that failure to control for measurement error, and confounding variables are likely to result in a negative bias in the effect of class-average achievement. The direction of this bias works against the BFLPE so that the controls make the BFLPE even more significant and more negative. In this sense, the BFLPE is robust concerning this bias. In contrast, failure to control measurement error and confounding variables produces a positive bias for the peer spillover effect on achievement. Thus, the direction of the bias is in the same direction as the prediction of a positive peer spillover effect. These contrasting effects and implications are apparent in our analysis. Hence, claims of positive peer spillover effects without these controls must be viewed cautiously. Furthermore, because there will always be unmeasured confounding variables likely to be positively related to class-average achievement, it might only be possible to resolve this problem partially. Still, the robustness of our findings revealed a negative BFLPE across all modeling specifications, supporting the characterization of the effect as a “pan-human universal” phenomenon (Marsh et al., 2020; Marsh, Pekrun, et al., 2018; Marsh, Xu et al., 2021).

### Replicating Findings of Previous Studies

Replications, being defined as intentional attempts to repeat previous research to verify or disprove the findings reported in the past (Plucker & Makel, 2021), are important for developing a “cumulative knowledge base” (Peterson & Schreiber, 2012, p. 287). Theoretical conclusions are stronger when they are based on this accumulated knowledge, and they can make a valuable and meaningful contribution to the development of educational policy and practice. Following Dicke et al.’s (Dicke et al., 2018; Televantou et al., 2021) methodological approach, we consider the impact of failing to account for measurement error on compositional effect estimates. Dicke et al. found that an under-specified model leads to less positive peer spillover effect, and more negative BFLPEs. We replicate their findings, enhancing the external validity of their claims. With doubly latent models and adjustments for student background, the initially detected peer spillover effect disappears altogether for USA and New Zealand data. For Canada (Ontario), the effect becomes substantially smaller — but it remains positive and statistically significant. In a supplementary study, we verified this finding using three different sets of student background variables to correct for selection bias at the student level (see Supplementary Materials, Table A.6). Whether and how selection bias can be sufficiently addressed in observational studies have been highly debated (Reardon & Owens, 2014) since no study can effectively control for the infinite number of potential confounders at the student level. By evaluating models based on different sets of student-level covariates and demonstrating the same trend in the findings, we address relevant concerns (Dicke et al., 2018).

### Resolving a Theoretical Paradox

The divergence in conclusions regarding the magnitude of the peer spillover effect, even after controls for measurement and omitted variable bias, echoes previous studies that

are also contradictory, with some supporting the existence of positive achievement compositional effects and some rejecting it (Becker et al., 2021; Sacerdote, 2011). Supplementary analysis to our main study (“Modeling Mathematics Self-concept as a Mediator”) demonstrates how class-average achievement can have different effects on student math self-concept and achievement, even though math self-concept and achievement are reciprocally related (see claims of Becker et al., 2021). It suggests that part of the effect of class-average achievement is mediated via mathematics self-concept, and this mediated effect is negative (Marsh, 2023; Marsh, Pekrun et al., 2023). Thus, another explanation of the theoretical paradox initially identified and partially explained by Dicke et al. (2018) might be derived: the BFLPE could be one mechanism driving negative compositional effects on achievement; however, other factors at the level of the classroom or the school (e.g., instructional practices) may also operate so that peer spillover effects are, eventually, manifested. Future research could address this hypothesis.

### Limitations and Directions for Future Research

Our study investigates the impact of failing to account for measurement error bias in compositional analysis with SIMS80 data by applying the doubly latent approach. Intact classes were used in the sampling of students in our study. Hence, we were concerned about whether overcorrecting for sampling error affected our estimates (see Marsh et al., 2012, for a relevant discussion). However, no substantial differences were observed in juxtaposing estimates of our analyses obtained with models assuming manifest aggregation (the latent manifest approach; Lüdtke et al., 2008; see Supplementary Materials, Table A.5).

We also note that the imputation model we used for missing data, despite considering all the multilevel covariates used in our analysis, does not mimic the analytical model used: implementing multilevel imputation would be ideal for our study. However, we faced serious convergence issues when we tried to do so.

In testing mediation, mathematics self-concept and achievement were measured at the end of the school year (see Fig. 2; dashed line) since prior measures of mathematics self-concept were not available with SIMS80. Thus, we base our mediational analysis on cross-sectional data, with both the mediator and the predictor measured at the end-of-year exam (for problems with mediation based on cross-sectional data, see Maxwell et al., 2011; O’Laughlin et al., 2018). While existing literature suggests that self-concept and individual achievement are reciprocally related (REM; Marsh & Craven, 2005; Marsh, 2023; Marsh, Pekrun et al., 2022), we only model the skill development process (i.e., prior achievement leads to subsequent academic self-concept). In this respect, our mediation analysis is more like a “what if” exploratory study of an interesting question, and our findings are only tentative, leading to a “hypothesis” that can be further tested when appropriate data are available to test mediation (e.g., a study with three or more waves of data). We note that the apparent lag 0 effect of academic self-concept on academic achievement might merely reflect lag 1 effects (effects of academic self-concept at a previous time point), not included in the model (Marsh, Pekrun et al., 2022). A promising direction for future research would be the application of models with both

reciprocal and contemporaneous effects between achievement and self-concept (Muthen & Asparouhov, 2023).

The strength of our analysis from a substantive point of view is that it shows a negative and statistically significant BFLPE, controlling for true measures of prior achievement — that persists and becomes even more prominent after adjustments for measurement error. There is now a vast literature in support of the BFLPE. Although most of this research is based on cross-sectional data, several studies have also evaluated it longitudinally. Interestingly, the results based on cross-sectional (e.g., Nagengast & Marsh, 2012;  $ES = -.286$ ), and longitudinal (e.g., Dicke et al., 2018;  $ES = -.36$ ) analyses do not differ substantially in the size of the effect. Our study replicates and extends these findings. What is interesting in putting the peer spillover effect and the BFLPE together in the same model is how the application of progressively stronger models results in systematically weaker (less positive, null, or even negative) peer spillover effects and systematically stronger (more negative) BFLPEs. Importantly, the direction of changes in both the spillover and BFLPE is consistent with a priori predictions.

Our estimates of compositional effects do not, however, represent an unambiguous causal effect; our interpretation would be strengthened by juxtaposing our results with potentially stronger designs such as regression discontinuity, propensity score matching (Randolph & Falbe, 2014), instrumental variables (e.g., Aral & Nicolaides, 2017), or a true experiment with random assignment (Paloyo, 2020). Existing studies have also pointed to the potential of social networks literature (e.g., Froehlich et al., 2020) in informing research on peer effects (Paloyo, 2020). Thus, for example, Koivuhovi et al. (2022) found that peer-group-average achievement had no effect on academic self-concept beyond the negative (BFLPE) effect of class-average achievement. All these are interesting avenues for exploration in future studies. Nevertheless, current research — including the present investigation — suggests that peer spillover effects are substantially smaller when appropriate adjustments are made, and may even disappear altogether.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s10648-023-09816-3>.

**Acknowledgements** Ioulia Televantou acknowledges Christos Nicolaides for providing funding and support through a starting grant awarded by the University of Cyprus in the first stages of the implementation of this research.

**Funding** Open access funding provided by the Cyprus Libraries Consortium (CLC).

**Data Availability** The data supporting the findings of this study are available online at: <https://www.gu.se/en/center-for-comparative-analysis-of-educational-achievement-compeat/studies-before-1995/second-international-mathematics-study-1980>. The data is provided in SPSS format and can be accessed without restrictions. For any inquiries about the data or access issues, please contact Ioulia Televantou at [i.televantou@euc.ac.cy](mailto:i.televantou@euc.ac.cy)

## Declarations

**Conflict of Interest** The authors declare no competing interests.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Ammermueller, A., & Pischke, J. S. (2009). Peer effects in European primary schools: Evidence from the progress in international reading literacy study. *Journal of Labor Economics*, 27(3), 315–348. <https://doi.org/10.1086/603650>
- Aral, S., & Nicolaides, C. (2017). Exercise contagion in a global social network. *Nature Communications*, 8, 14753. <https://doi.org/10.1038/ncomms14753>
- Basarkod, G., Marsh, H. W., Guo, J., Dicke, T., Xu, K., & Parker, P. D. (2023). The Big-Fish Little-Pond Effect for reading self-beliefs: A crossnational exploration with PISA 2018. *Scientific Studies of Reading*, 27(4), 375–392. <https://doi.org/10.1080/10888438.2023.2174028>
- Becker, M., Kocaj, A., Jansen, M., Dumont, H., & Lüdtke, O. (2021). Class-average achievement and individual achievement development: Testing achievement composition and peer spillover effects using five German longitudinal studies. *Journal of Educational Psychology*, 114(1), 177–197. <https://doi.org/10.1037/edu0000519>
- Brown, M. (1996). FIMS and SIMS: the first two IEA International Mathematics Surveys. *Assessment in Education: Principles, Policy & Practice*, 3(2), 193–212. <https://doi.org/10.1080/0969594960030206>
- Burns, R., & Mason, D. (2002). Class composition and student achievement in elementary schools. *American Educational Research Journal*, 39(1), 207–233. <https://doi.org/10.3102/00028312039001207>
- Byrne, B. M., & Shavelson, R. J. (1986). On the structure of adolescent self-concept. *Journal of Adolescent Psychology*, 78(6), 474–481. <https://doi.org/10.1037/0022-0663.78.6.474>
- Caro, D. H., Kyriakides, L., & Televantou, I. (2017). Addressing omitted prior achievement bias in international assessments: An applied example using PIRLS-NPD matched data. *Assessment in Education: Principles, Policy & Practice*, 25(1), 5–27. <https://doi.org/10.1080/0969594X.2017.1353950>
- Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & Robert, L. (1966). *Equality of educational opportunity*. US Government Printing Office.
- Cooley Fruehwirth, J. (2013). Identifying peer achievement spillovers: Implications for desegregation and the achievement gap. *Quantitative Economics*, 4, 85–124. <https://doi.org/10.3982/QE93>
- Colom, R., & Flores-Mendoza, C. (2007). Intelligence Predicts Scholastic Achievement Irrespective of SES Factors: Evidence from Brazil. *Intelligence*, 35, 243–251. <https://doi.org/10.1016/j.intell.2006.07.008>
- Dicke, T., Marsh, H. W., Parker, P. D., Pekrun, R., Guo, J., & Televantou, I. (2018). Effects of school-average achievement on individual self-concept and achievement: Unmasking phantom effects masquerading as true compositional effects. *Journal of Educational Psychology*, 110(8), 1112–1126. <https://doi.org/10.1037/edu0000259P>
- Dishion, T. J., & Tipsord, J. M. (2011). Peer contagion in child and adolescent social and emotional development. *Annual Review of Psychology*, 62(1), 189–214. <https://doi.org/10.1146/annurev.psych.093008.100412>
- Epple, D., & Romano, R. E. (2011). Peer effects in education: A survey of the theory and evidence. In *Handbook of social economics* (Vol. 1, pp. 1053–1163). <https://doi.org/10.1016/B978-0-444-53707-2.00003-7>
- Fauth, B., Atlay, C., Dumont, H., & Decristan, J. (2021). Does what you get depend on who you are with? Effects of student composition on teaching quality. *Learning and Instruction*, 71, 101355. <https://doi.org/10.1016/j.learninstruc.2020.101355>
- Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, 7(2), 117–140.
- Fleischmann, M., Huebner, N., Marsh, H. W., Trautwein, U., & Nagengast, B. (2021). Investigating the association between the big fish little pond effect and grading on a curve: A large-scale quasi-experimental

- study. *International Journal of Educational Research*, 110, 101853. <https://doi.org/10.1016/j.ijer.2021.101853>. Get rights and content
- Froehlich, D. E., Van Waes, S., & Schäfer, H. (2020). Linking quantitative and qualitative network approaches: A review of mixed methods social network analysis in education research. *Review of Research in Education*, 44(1), 244–268. <https://doi.org/10.3102/0091732X20903311>
- Guay, F., Larose, S., & Boivin, M. (2004). Academic self-concept and educational attainment level: A ten-year longitudinal study. *Self and Identity*, 3(1), 53–68. <https://doi.org/10.1080/13576500342000040>
- Guo, J., Marsh, H. W., Parker, P. D., Dicke, T., & Van Zanden, B. (2018). Cross-cultural generalizability of social and dimensional comparison effects on reading, math, and science self-concepts for primary school students using the combined PIRLS and TIMSS data. *Learning and Instruction*, 58, 210–219. <https://doi.org/10.1016/j.learninstruc.2018.07.007>
- Hanushek, E. A., Kain, J. F., Markman, J. M., & Rivkin, S. G. (2003). Does peer ability affect student achievement? *Journal of Applied Econometrics*, 18(5), 527–544. <https://doi.org/10.1002/jae.741>
- Hanushek, E. A., & Woessmann, L. (2011). The economics of international differences in educational achievement. *Handbook of the Economics of Education*, 3, 89–200.
- Harker, R., & Tymms, P. (2004). The effects of student composition on school outcomes. *School Effectiveness and School Improvement*, 15(2), 177–199. <https://doi.org/10.1076/sesi.15.2.177.30432>
- Hutchison, D. (2004). The effect of measurement errors on apparent group-level effects in educational progress. *Quality and Quantity*, 38(4), 407–424. <https://doi.org/10.1023/B:QUQU.0000043116.22582.fd>
- Hutchison, D. (2007). When is a compositional effect not a compositional effect? *Quality and Quantity*, 41(2), 219–232. <https://doi.org/10.1007/s11135-007-9094-2>
- Ireson, J., Hallam, S., & Plewis, I. (2001). Ability grouping in secondary schools: Effects on pupils' self-concepts. *British Journal of Educational Psychology*, 71(2), 315–326. <https://doi.org/10.1348/000709901158541>
- Ivanova, M., & Michaelides, M. P. (2022). Motivational components in TIMSS 2015 and their effects on engaging teaching practices and mathematics performance. *Studies in Educational Evaluation*, 74, 101173. <https://doi.org/10.1016/j.stueduc.2022.101173>
- Koivuhovi, S., Marsh, H. W., Dicke, T., Sahdra, B., Guo, J., Parker, P. D., & Vainikainen, M.-P. (2022). Academic self-concept formation and peer-group contagion: Development of the big-fish-little-pond effect in primary-school classrooms and peer groups. *Journal of Educational Psychology*, 114(1), 198–213. <https://doi.org/10.1037/edu0000554>
- Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling*, 9(2), 151–173. [https://doi.org/10.1207/S15328007SEM0902\\_1](https://doi.org/10.1207/S15328007SEM0902_1)
- Little, T. D., Rhemtulla, M., Gibson, K., & Schoemann, A. M. (2013). Why the items versus parcels controversy needn't be one. *Psychological Methods*, 18, 285–300. <https://doi.org/10.1037/a0033266>
- Little, T. D., Rioux, C., Odejimi, O. A., & Stickley, Z. L. (2022). Parceling in structural equation modeling: A comprehensive introduction for developmental scientists. *Elements in Research Methods for Developmental Science*. <https://doi.org/10.1017/9781009211659>
- Loyalka, P., Zakharov, A., & Kuzmina, Y. (2018). Catching the big fish in the little pond effect: Evidence from 33 countries and regions. *Comparative Education Review*, 62(4), 542–564. <https://doi.org/10.1086/699672>
- Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multi-level latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, 13(3), 203. <https://doi.org/10.1037/a0012869>
- Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies*, 60(3), 531–542. <https://doi.org/10.2307/2298123>
- Marsh, H. W. (2023). Extending the reciprocal effects model of math self-concept and achievement: Long-term implications for end-of-high-school age-26 outcomes and long term expectations. *Journal of Educational Psychology*, 115(2), 193–211. <https://doi.org/10.1037/edu000075010.1037/edu0000750>
- Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology*, 79(3), 280–295 <http://files.eric.ed.gov/fulltext/ED278685.pdf>
- Marsh, H. W., Abduljabbar, A. S., Morin, A. J., Parker, P., Abdelfattah, F., Nagengast, B., & Abu-Hilal, M. M. (2015). The big-fish-little-pond effect: Generalizability of social comparison processes over two age cohorts from Western, Asian, and Middle Eastern Islamic countries. *Journal of Educational Psychology*, 107(1), 258.
- Marsh, H. W., & Craven, R. G. (2005). A reciprocal effects model of the causal ordering of self-concept and achievement: New support for the benefits of enhancing self-concept. In H. W. Marsh, R. G. Craven,


- & D. M. McInerney (Eds.), *International advances in self research: The new frontiers of self-research* (Vol. 2, pp. 15–51). Information Age.
- Marsh, H. W., & 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), 133–163. <https://doi.org/10.1111/j.1745-6916.2006.00010.x>
- Marsh, H. W., & 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, 364–376. <https://doi.org/10.1037/0003-066X.58.5.364>
- Marsh, H. W., Kong, C. K., & 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), 337–349. <https://doi.org/10.1037/0022-3514.78.2.337>
- Marsh, H. W., Kuyper, H., Morin, A. J. S., Parker, P. D., & Seaton, M. (2014). Big–fish–little–pond social comparison and local dominance effects: Integrating new statistical models, methodology, design, theory and substantive implications. *Learning and Instruction*, 33, 50–66. <https://doi.org/10.1016/j.learninstruc.2014.04.002>
- Marsh, H. W., Lüdtke, O., Nagengast, B., Morin, A. J., & 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(3), 257–284. <https://doi.org/10.1037/a0032773>
- Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J., Abduljabbar, A. S., & Köller, O. (2012). Classroom climate and contextual effects: Conceptual and methodological issues in the evaluation of group–level effects. *Educational Psychologist*, 47(2), 106–124. <https://doi.org/10.1080/00461520.2012.670488>
- Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B., & 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), 764–802. <https://doi.org/10.1080/00273170903333665>
- Marsh, H. W., & Martin, A. J. (2011). Academic self–concept and academic achievement: Relations and causal ordering. *British Journal of Educational Psychology*, 81(1), 59–77.
- Marsh, H. W., Parker, P. D., Guo, J., Pekrun, R., & Basarkod, G. (2020). Psychological Comparison Processes and Self–Concept in Relation to Five Distinct Frame–Of–Reference Effects: Pan–Human Cross–Cultural Generalizability over 68 Countries. *European Journal of Personality*, 34(2), 180–202. <https://doi.org/10.1002/per.2232>
- Marsh, H. W., Pekrun, R., Dicke, T., Guo, J., Parker, P. D., & Basarkod, G. (2023). Disentangling the long–term compositional effects of schoolaverage achievement and SES: A substantive–methodological synergy. *Educational Psychology Review*, 35(3), 70. <https://doi.org/10.1007/s10648-023-09726-4>
- Marsh, H. W., Pekrun, R., & Lüdtke, O. (2022). Directional ordering of self–concept, school grades, and standardized tests over five years: New tripartite models juxtaposing within and between–person perspectives. *Educational Psychology Review*, 34(4), 2697–2744. <https://doi.org/10.1007/s10648-022-09662-9>
- Marsh, H. W., Pekrun, R., Murayama, K., Arens, A. K., Parker, P. D., Guo, J., & Dicke, T. (2018). An integrated model of academic self–concept development: Academic self–concept, grades, test scores, and tracking over 6 years. *Developmental Psychology*, 54(2), 263–280. <https://doi.org/10.1037/dev0000393>
- Marsh, H. W., Seaton, M., Kuyper, H., Dumas, F., Huguet, P., Régner, I., Buunk, A. P., Monteil, J. M., & Gibbons, F. X. (2010). Phantom behavioral assimilation effects: Systematic biases in social comparison choice studies. *Journal of Personality*, 78(2), 671–710. <https://doi.org/10.1111/j.1467-6494.2010.00630.x>
- Marsh, H. W., Seaton, M., Trautwein, U., Lüdtke, O., Hau, K. T., O’Mara, A. J., & 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, 319–350. <https://doi.org/10.1007/s10648-008-9075-6>
- Marsh, H., Xu, M., Parker, P., Hau, K. T., Pekrun, R., Elliot, A., Guo, J., Dicke, T., & Basarkod, G. (2021). Moderation of the Big–Fish–Little–Pond Effect: Juxtaposition of Evolutionary (Darwinian–Economic) and Achievement Motivation Theory Predictions Based on a Delphi Approach. *Educational Psychology Review*, 33(4), 1353–1378. <https://doi.org/10.1007/s10648-020-09583-5>

- Maxwell, S. E., Cole, D. A., & Mitchell, M. A. (2011). Bias in cross-sectional analyses of longitudinal mediation: Partial and complete mediation under an autoregressive model. *Multivariate Behavioral Research*, *46*(5), 816–841. <https://doi.org/10.1080/00273171.2011.606716>
- Mortimore, P., Sammons, P., Stoll, L., & Ecob, R. (1988). *School matters*. Univ of California Press.
- Muthén, B., & Asparouhov, T. (2023). Can cross-lagged panel modeling be relied on to establish cross-lagged effects? The case of contemporaneous and reciprocal effects. Retrieved from: <https://www.statmodel.com/download/ReciprocalV3.pdf>
- Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8th ed.). Authors.
- Nagengast, B., & 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*(4), 1033.
- Nomi, T., & Raudenbush, S. W. (2016). Making a success of “Algebra for All”: The impact of extended instructional time and classroom peer skill in Chicago. *Educational Evaluation and Policy Analysis*, *38*(2), 431–451. <https://doi.org/10.3102/0162373716643756>
- Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., et al. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, *73*, 719–748.
- O’Laughlin, K. D., Martin, M. J., & Ferrer, E. (2018). Cross-sectional analysis of longitudinal mediation processes. *Multivariate Behavioral Research*, *53*(3), 375–402. <https://doi.org/10.1080/00273171.2018.1454822>
- Opendakker, M.-C., Van Damme, J., De Fraine, B., Van Landeghem, G., & Onghena, P. (2002). The effect of schools and classes on mathematics achievement. *School Effectiveness and School Improvement*, *13*(4), 399–427. <https://doi.org/10.1076/sesi.13.4.399.10283>
- Paloyo, A. R. (2020). Peer effects in education: recent empirical evidence. In *The Economics of Education* (pp. 291–305). Academic Press. <https://doi.org/10.1016/B978-0-12-815391-8.00021-5>
- Pearl, J. (2002). Causal inference in the health sciences: A conceptual introduction. *Health Services and Outcomes Research Methodology*, *2*, 189–220. <https://doi.org/10.1023/A:1020315127304>
- Peterson, S. E., & Schreiber, J. B. (2012). Personal and interpersonal motivation for group projects: Replications of an attributional analysis. *Educational Psychology Review*, *24*(2), 287–311. <https://doi.org/10.1007/s10648-012-9193-z>
- Plucker, J. A., & Makel, M. C. (2021). Replication is important for educational psychology: Recent developments and key issues. *Educational Psychologist*, *56*(2), 90–100. <https://doi.org/10.1080/00461520.2021.1895796>
- Pokropek, A. (2015). Phantom effects in multilevel compositional analysis: Problems and solutions. *Sociological Methods & Research*, *44*(4), 677–705. <https://doi.org/10.1177/0049124114553801>
- Randolph, J. J., & Falbe, K. (2014). A step-by-step guide to propensity score matching in R. *Practical Assessment, Research & Evaluation*, *19*(18) Available online: <http://pareonline.net/getvn.asp?v=19&n=18>
- Raykov, T., Marcoulides, G. A., & Millsap, R. E. (2013). Factorial invariance in multiple populations: A multiple testing procedure. *Educational and Psychological Measurement*, *73*(4), 713–727. <https://doi.org/10.1177/0013164412451978>
- Reardon, S. F., & Owens, A. (2014). 60 years after brown: Trends and consequences of school segregation. *Annual Review of Sociology*, *40*(1), 199–218. <https://doi.org/10.1146/annurev-soc-071913-043152>
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley.
- Sacerdote, B. (2001). Peer effects with random assignment: Results for Dartmouth roommates. *The Quarterly Journal of Economics*, *116*(2), 681–704.
- Sacerdote, B. (2011). Peer effects in education: How might they work, how big are they and how much do we know thus far? In *Handbook of the Economics of Education* (Vol. 3, pp. 249–277). Elsevier. <https://doi.org/10.1016/B978-0-444-53429-3.00004-1>
- Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, *7*(2), 147–177. <https://doi.org/10.1037/1082-989X.7.2.147>
- Seaton, M., Marsh, H. W., & Craven, R. G. (2009). Earning its place as a pan-human theory: Universality of the big-fish-little-pond effect across 41 culturally diverse countries. *Journal of Educational Psychology*, *101*, 403–419. <https://doi.org/10.1037/a0013838>
- Snijders, T. A. & Bosker, RJ (2012). Multilevel analysis: An introduction to basic and advanced multilevel modelling.
- Stäbler, F., Dumont, H., Becker, M., & Baumert, J. (2017). What happens to the fish’s achievement in a little pond? A simultaneous analysis of class-average achievement effects on achievement and academic



- self-concept. *Journal of Educational Psychology*, 109(2), 191–207. <https://doi.org/10.1037/edu000135>
- Stinebrickner, R., & Stinebrickner, T. R. (2006). What can be learned about peer effects using college roommates? Evidence from new survey data and students from disadvantaged backgrounds. *Journal of Public Economics*, 90(8-9), 1435–1454. <https://doi.org/10.1016/j.jpubeco.2006.03.002>
- Teddlie, C., Stringfield, S. & 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., Dicke, T., & Nicolaides, C. (2021). Phantom and big-fish-little-pond-effects on academic self-concept and academic achievement: Evidence from English early primary schools. *Learning and Instruction*, 71, 101399. <https://doi.org/10.1016/j.learninstruc.2020.101399>
- Televantou, I., Marsh, H. W., Kyriakides, L., Nagengast, B., Fletcher, J., & Malmberg, L. E. (2015). 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*, 26(1), 75–101. <https://doi.org/10.1080/09243453.2013.871302>
- Thrupp, M. (1999). *Schools making a difference: school mix, school effectiveness, and the social limits of reform*. McGraw-Hill Education (UK).
- Thrupp, M., Lauder, H., & Robinson, T. (2002). School composition and peer effects. *International Journal of Educational Research*, 37(5), 483–504. [https://doi.org/10.1016/S0883-0355\(03\)00016-8](https://doi.org/10.1016/S0883-0355(03)00016-8)
- Van de Werfhorst, H. G., & Mijs, J. J. B. (2010). *Annual Review of Sociology*, 36, 407–428. <https://doi.org/10.1146/annurev.soc.012809.102538>
- Wagner, G. (2022). How group composition affects gifted students: theory and evidence from school effectiveness studies. *Gifted and Talented International*, 37(1), 1–13. <https://doi.org/10.1080/15332276.2021.1951145>
- West, S. G., & Thoemmes, F. (2010). Campbell's and Rubin's perspectives on causal inference. *Psychological Methods*, 15(1), 18–37. <https://doi.org/10.1037/a0015917>
- Willms, J. D. (1985). The balance thesis: contextual effects of ability on pupils' O-grade examination results. *Oxford Review of Education*, 11(1), 33–41. <https://doi.org/10.1080/0305498850110103>
- Woodhouse, G., Yang, M., Goldstein, H., & Rasbash, J. (1996). Adjusting for measurement error in multilevel analysis. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 159(2), 201–212. <https://doi.org/10.2307/2983168>
- Yang Hansen, K., Radišić, J., Ding, Y., & Liu, X. (2022). Contextual effects on students' achievement and academic self-concept in the Nordic and Chinese educational systems. *Large-scale Assessments in Education*, 10(1), 1–26. <https://doi.org/10.1186/s40536-022-00133-9>
- Zimmer, R. W., & Toma, E. F. (2000). Peer effects in private and public schools across countries. *Journal of Policy Analysis and Management: The Journal of the Association for Public Policy Analysis and Management*, 19(1), 75–92.
- Zirkel, P. A. (1971). Self-concept and the “disadvantage” of ethnic group membership and mixture. *Review of Educational Research*, 41, 211–225 <http://www.jstor.org/stable/116952>

## Authors and Affiliations

Ioulia Televantou<sup>1,2</sup>  · Herbert W. Marsh<sup>3,4</sup> · Kate M. Xu<sup>5</sup> · Jiesi Guo<sup>3</sup> · Theresa Dicke<sup>3</sup>

✉ Ioulia Televantou  
i.televantou@euc.ac.cy

Herbert W. Marsh  
Herb.Marsh@acu.edu.au

Kate M. Xu  
kate.xu@ou.nl

Jiesi Guo  
Jiesi.Guo@acu.edu.au

Theresa Dicke  
Theresa.Dicke@acu.edu.au

<sup>1</sup> Department of Education Sciences, European University Cyprus, 6 Diogenis Str., 2404 Engkomi, P.O. Box: 22006, 1516 Nicosia, Cyprus

<sup>2</sup> Theological School of the Church of Cyprus, Isokratous 1-7, 1016 Nicosia, Cyprus

<sup>3</sup> Institute for Positive Psychology and Education (IPPE), Australian Catholic University, North Sydney, NSW 20160 2135, Australia

<sup>4</sup> Department of Education, Oxford University, 15 Norham Gardens, Oxford OX2 6PY, UK

<sup>5</sup> Faculty of Educational Sciences, Open University of the Netherlands, 6419, AT, Heerlen, The Netherlands