Effects of School-Average Achievement on Individual Self-concept and Achievement: Unmasking Phantom Effects Masquerading as True Compositional Effects

Theresa Dicke[a], Herbert W. Marsh[a], Philip D. Parker[a], Reinhard Pekrun[b], Jiesi Guo[a], & Ioulia Televantou[c]

[a] Australian Catholic University, Institute for Positive Psychology and Education, Locked Bag 2002, Strathfield NSW 2135, Australia; Theresa.dicke@acu.edu.au; Herb.Marsh@acu.edu.au; Philip.Parker@acu.edu.au; Jiesi.Guo@acu.edu.au

[b] Ludwig-Maximilian University, Munich, Department of Psychology, Leopoldstr. 13, 80802 München, Germany; pekrun@lmu.de

[c] University of Cyprus, Nicosia, Cyprus, Ioulia.Televantou@gmail.com

Author note:
Theresa Dicke,
Institute for Positive Psychology and Education,
Australian Catholic University,
Level 9, 33 Berry Street, North Sydney, NSW 2060

Australia

email: theresa.dicke@acu.edu.au

Effects of School-Average Achievement on Individual Self-concept and Achievement: Unmasking Phantom Effects Masquerading as True Compositional Effects

Abstract

School-average achievement is often reported to have positive effects on individual achievement (peer spillover effect). However, it is well established that school-average achievement has negative effects on academic self-concept (big-fish-little-pond effect; BFLPE) and that academic self-concept and achievement are positively correlated and mutually reinforcing (reciprocal effects model; REM). We resolve this theoretical paradox based on a large, longitudinal sample (N=14,985 US children) and improved methodology. More appropriate multilevel modeling that controls for phantom effects (due to measurement error and pre-existing differences) make the BFLPE even more negative, but turn the peer spillover effect from positive to slightly below zero. Thus, attending a high-achieving school has negative effects on academic self-concept and a non-positive effect on achievement. The results question previous studies and meta-analyses showing a positive peer spillover effect that do not control for phantom effects, along with previous policy and school selection decisions based on this research.

Educational Impact And Implications Statement

Counter to a widely held belief that being in a high-achieving school has a positive effect on student's achievement (peer-spillover effect), the present findings suggest that this effect is actually slightly negative. When using stronger, more appropriate statistical methodology, the apparent peer- spillover effect disappeared, suggesting that positive effects are a phantom. Furthermore, the negative effect of school-average achievement on academic self-concept ("big-fish-little-pond effect") turned out to be even more negative when using more appropriate methodology. Thus, the findings indicate that attending a high-achieving school

has a negative effect on self-concept and no positive effect on achievement. These results call into question prior research that did not control for phantom effects, and challenge policy and practice decisions that promote selective schooling.

Keywords: phantom effects, academic self-concept, big-fish-little-pond effect, peer spillover effect, academic equality

Effects of School-Average Achievement on Individual Self-concept and Achievement: Unmasking Phantom Effects Masquerading as True Compositional Effects

Think back to when you were a school student. When you performed well, your beliefs in your own abilities most likely were boosted; when you believed in yourself, you probably performed better. But of course you're not alone in school; you are surrounded by school- and classmates. So now consider the impact of your fellow students. Imagine being in a high-achieving school, how might this context affect the beliefs in your own abilities and your actual achievement? How do these effects change if you imagine being in a low-achieving school?

According to research these questions are addressed by *school composition effects* (Harker & Tymms , 2004; Marsh et al., 2009;Marsh et al., 2012; Willms, 1985[1]). These are defined to be present whenever a given predictor variable at the aggregated level (e.g., school level) has an effect on an individual outcome variable, over and beyond the effect of the same predictor variable at the individual level (Harker & Tymms, 2004; Nash, 2003). Two of the most prominent school compositional effects in educational psychology are the Big fish little pond effect (BFLPE; Marsh, Seaton, et al., 2008) and the peer spillover effect (Willms, 1985).

The BFLPE implies that aggregated achievement negatively predicts self-beliefs such as academic self-concept (ASC; e.g., Marsh, Seaton, et al., 2008) over and above the positive effect of individual achievement on the same variables at the individual level. Thus, according to the BFLPE a student with a given level of achievement will have a lower ASC in a high achieving (big pond) class than in a low achieving (little pond) class (see Figure 1a;.e.g., Marsh, Seaton, et al., 2008).

The peer spillover effect (Cooley Fruehwirth, 2013) however, implies a positive effect of aggregated achievement on subsequent individual achievement, over and above the positive effect of prior individual achievement at the individual level (see Figure 1b; e.g., De Fraine, Van Damme, Van Landeghem, Opdenakker, & Onghena, 2003; Stäbler, Dumont,

Becker, & Baumert, 2016; Willms, 1985). Hence, according to this research the student's achievement will be higher in a high achieving class than in a low achieving class.

Taken together, these findings imply that attending a school or class with high achievers has a negative impact on a student's ASC, but positively impacts a student's individual achievement. When considering the proposed positive reciprocal relationship of ASC and academic achievement (REM) however (see Figure 1c), it becomes obvious that these are apparently paradoxical effects (see Figure 1d and Figure 1).

A major clue for resolving this theoretical paradox might lie in the validity of the statistical models used to support these seemingly contradictory conclusions. While increasingly sophisticated doubly-latent multilevel models have been at the forefront of BFLPE research (Marsh, et al, 2009), the vast majority of studies in support of peer spillover effects are based on traditional (manifest) multilevel models with poor controls. This is important because, as we will show, stronger controls of measurement error and pre-existing differences (phantom effects) through, for example, the use of doubly-latent models, are likely to results in a more negative BFLPE, but a less positive peer spillover effect that might even go from positive to negative. Indeed, Harker and Tymms (2004) first coined the term "phantom effect" based on their findings that with appropriate statistical models, positive composition effects disappeared – now you see it, now you do not.

While there is strong empirical evidence and a large number of studies to support both the BFLPE (see Marsh & Seaton, 2015 for an overview) and the REM (Marsh & Craven, 2006), results regarding the peer spillover effect, although widely accepted, are inconsistent (Hattie, 2002; Hutchinson, 2007; Nash, 2003; Televantou et al., 2015). Moreover, methodological advances show that the presence of such positive compositional effects of achievement might be the result of unreliability at the individual level for which has not been accounted. Consequently, this lack of consideration of unreliability causes bias on effects at the group level after aggregation (Hutchinson, 2007; Marsh, Seaton et al., 2008 Televantou et

al., 2015). Research in the area of predicting educational outcomes that does not utilize random assignment to conditions is additionally problematic as it is never possible to consider 100% of the pre-existing differences (Goldring,1990; Harker & Tymms, 2004, Strand, 2010; Marks, 2015). Importantly, residual variance associated with pre-existing differences will almost always bias the results in favor of students in high-ability groups, even in longitudinal studies with reasonable pre-treatment variables (Craven & Marsh, 2000). Overall, such limitations need to be considered not only in research but moreover with regard to debates on educational policies (Goldring ,1990; Thrupp, Lauder, & Robinson, 2002; Pokropek, 2015).

These inaccuracies of analysis and model specification can be subsumed under the label *phantom effect* (Harker & Tymms, 2004; Televantou et al., 2015). For the present study, we have a detailed look at two different sources of phantom effects: a) measurement error at the individual level that leads to positively biased estimates of the compositional effects; and b) the lack of including sufficient control for pre-existing differences, which can also lead to the appearance of exaggerated cross-level effects that are falsely interpreted as compositional effects.

Hence, in the present study we test the REM, BFLPE, and the peer spillover effect in a longitudinal sample of students in the U.S. (grades 1, 3, and 5). First, we test the mere appearance of the REM and both compositional effects in a simple baseline model. Second, we test their persistence after including different approaches to minimize the likelihood of the Phantom Effect in our model. Thus, we a) correct for measurement error in our achievement measures; b) include prior achievement at kindergarten as an additional indicator for the baseline measure of achievement; and c) include several covariates that also contribute to school composition, such as gender, ethnic heritage, and socioeconomic status (SES), to ensure a less positively biased compositional effect of average achievement. Revealing the peer spillover effect as a phantom would call into question all previous studies showing a positive peer spillover effect. Moreover, there would be serious implications for policies and

decisions based on this research that does not control for a phantom effect.

The important role of self-concept in the academic context – the reciprocal effects model

While it seems natural to look at subsequent achievement as an outcome in educational research, ASC, the perception of one's own abilities (Marsh & Shavelson, 1985), has also emerged as a positive, important educational outcome within the past years (Marsh, 2007). Robust evidence exists for the first of the above mentioned processes. Indeed, there seem to be two complementary processes: the skill enhancement (i.e., ASC causes achievement) and skill development process (i.e., achievement causes ASC; proposed by Calsyn & Kenny, 1977) that jointly describe the relationship of ASC and achievement. Research has found support for both of these models, which indicates a reciprocal relationship of ASC and achievement (Reciprocal Effects Model; see Figure 1c; see e.g., Marsh & Craven, 2006 for an overview; Marsh & O'Mara, 2008). Furthermore two meta-analyses by Valentine, DuBois, and Cooper (2004) and Huang (2011) showed strong empirical evidence for this reciprocal relationship to be valid cross-culturally, as it generalized across age groups, gender, school-type, ethnicity, domains, and countries (for an overview see also Seaton, Marsh, Parker, Craven, & Yeung., 2015).

Compositional effects on academic achievement and self-beliefs

Generally the literature on school effectiveness distinguishes three types of effects: the actual effects of school processes and practices, structural effects, and compositional effects (see Marks, 2010). School processes and practices reflect the "pure" effect of schooling (e.g., classroom instructions). Structural effects refer to aspects such as school resources or facilities, or school size. School composition effects refer to the composition of (mainly) students (or peers) the school is made of. Studies often focus on the composition of SES (for an overview see e.g. Marks, 2015), gender and race (e.g., Strand, 2010), or academic achievement (e.g., Hutchinson, 2007; Willms, 1985). Thus, the occurrence and validity of compositional effects has very important practical implications as they also have a direct

impact on educational policies. They determine where the major responsibility for student achievement lies, such as with the school having a direct influence (student achievement can be strongly influenced by instruction practices and school processes), or an indirect influence through the school determining its own compositional structure (Thrupp et al., 2002). Two of the most prominent school compositional effects in educational research are the big-fish-little-pond effect (BFLPE; Marsh et al., 2008) and the peer spillover effect (Willms, 1985).

The Big-Fish-Little-Pond Effect. The BFLPE theorizes that students compare their own academic achievement with the achievements of their classmates (group level), and consequently this social comparison influences their ASC (e.g., Marsh, Seaton, et al., 2008; Nagengast & Marsh, 2012). With regard to the resulting interplay of group and individual level effects, the BFLPE thus hypothesizes that individual ability is positively related to ASC (i.e., the brighter I am the higher my ASC) but school-average achievement has a negative effect on ASC (see Figure 1b). Put simply, the brighter my classmates, the lower my ASC and vice versa. Extensive support for the BFLPE generalizes across student groups, subject domains, ASC instruments, and cultures (Marsh, Seaton et al., 2008). Thus, research conducted with three successive PISA data collections (Marsh & Hau, 2003: 103,558 students from 26 countries; Seaton, Marsh & Craven, 2010: 265,180 students from 41 countries; Nagengast & Marsh, 2012: 397,500 students from 57 countries) could show significant BFLPEs in 114 samples (overall it was present in 122 of 123 samples; see also Marsh et al. 2016).

Experimental studies have additionally provided strong evidence for the BFLPE (Zell & Alike, 2009, 2010). In five studies Zell and Alike (2009) showed the local dominance effect, which implies that provided with information of more or less local (peers within your immediate group, i.e., friends or classmates) or general/global (peers within a broader group, i.e., school mates) comparison information, individuals will use the most local level of comparison information provided to them. Thus, Zell and Alike (2009) provided participants

with feedback on a verbal reasoning test which had been completed by the participants earlier. The feedback was manipulated with regard to the feedback source (intragroup vs. intergroup comparisons) and the level of closeness of the feedback group (local vs. general). Results revealed that, when presented with the feedback, participants would use an intragroup and most local level available as a frame of reference, and clearly showed a big fish little pond effect. When intergroup and more general comparisons were presented in isolation, participants would also show a big fish little pond effect in those conditions.

Research on the longitudinal development of the BFLPE, however, is still scarce (Becker & Neumann, 2016). However, some studies could show that the BFLPE increases in primary school (Televantou, 2014), during high school (Marsh, Köller, & Baumert, 2001) and moreover this growing effect seems to be persistent up to four years after graduating high school (Marsh & O'Mara, 2010; Marsh, Trautwein, Lüdtke, Baumert, & Köller, 2007).

The Peer spillover Effect: A Phantom Effect? Over the past few decades some researchers have repeatedly claimed to have found a positive effect of school level achievement on subsequent individual achievement (see Figure 1b; e.g., De Fraine et al., 2003; Stäbler, Dumont, Becker, & Baumert, 2016; Willms, 1985). Hence, according to this research a student's achievement will be higher in a high achieving class than in a low achieving class. Taken together, these findings seem to imply that attending a school with high achievers reduces a student's ASC, but improves a student's individual achievement. However, as shown above, extensive research with the reciprocal effects model (Marsh & Craven, 2006) also shows that ASC and achievement are positively correlated and reciprocally related over time, resulting in a theoretical paradox (see Figure 1d and Figure 1).

This theoretical paradox might be resolved by testing the validity of statistical models used to support these seemingly contradictory conclusions. While there seems to be extensive research in favor of the BFLPE, results regarding the effects of group level achievement on individual achievement are still inconsistent (Nash, 2003). Specifically, other researchers have

only found very weak if any evidence for such a positive effect (e.g., Marks, 2010), or even a negative effect of group achievement on individual achievement (for an overview see Nash, 2003; Strand, 2010, Televantou et al., 2015). For example, Marsh (1991) evaluated the effect of school-average ability on the major outcomes of education. He used the large, nationally representative, and longitudinal High School and Beyond United States database, consisting of responses by the same high school students during grade 10 (T1), grade 12 (seniors; T2), and two years after the normal graduation from high school (T3). Among the components of the path analysis were: (a) individual-student and school-average measures of academic ability (a standardized test battery) and SES; (b) ASC, school grades, and educational aspirations measured at T1 and T2; and (c) university attendance, educational aspirations, and occupational aspirations at T3. The effects of school-average ability were negative for almost all of the grade 10, grade 12 and post-secondary outcomes; 16 of 18 effects were significantly negative and 2 were not statistically significant. Even though it might be argued that most of the important outcome variables in educational research were included in the database, the effect of school-average ability was not positive for a single outcome, most importantly for the present study being small but significantly negative for achievement and grades. For many grade 12 and post-secondary outcomes, there were: statistically significant negative effects of school-average ability even after controlling the substantial negative effects in grade 10 outcomes; and new, additional negative effects of school-average ability during the last two years of secondary school beyond the already substantial negative effects found early in secondary school. Many of the negative outcomes of school-average ability on many outcomes could be explained at least in part in terms of the negative effect of school-average-achievement on ASC—the BFLPE.

Additionally, several researchers have expressed their doubt with regard to the validity of the significant positive effect of school composition (e.g. Harker & Tymms, 2004; Hutchinson, 2007; Marks, 2015; Televantou et al., 2015; Thrupp et al., 2002). They not only

claim that untangling the effects of compositional effects (peers), teacher effects (e.g., instructional effectiveness), and school processes proves to be very complex (Marsh, Nagengast, Fletcher, & Televantou, 2011), but attribute the positive school composition effect to be a statistical artefact (Marks, 2015) or a so called phantom effect (Harker & Tymms, 2004; Televantou et al 2015; Nash, 2003).

Illustrating this sort of problem, in one of the strongest early meta-analyses of this literature, Goldring (1990) found that gifted students in special (homogeneous) classes were reported to have better academic achievement than their gifted counterparts in regular (heterogeneous) classes, but there were no systematic effects for self-concept. Goldring, however, emphasized that the magnitude of the effects in this area were questionable, as only one of the 24 studies used true assignment, while the other studies used weak matching procedures for distinguishing students in special and regular classes. Thus, Goldring found that the effects varied depending on different matching procedures, which has important implications for interpreting previous research. Thereby, she found the largest effects for achievement outcomes in studies where all non-equivalent group differences were assumed to be controlled by only one pretest variable, such as IQ, Race, or school achievement. When studies reported to control for more pretest variables the positive effect of being in a special class was largely reduced. Furthermore, in the only study that used true random assignment, the advantage of special classes disappeared altogether.

Moreover, Marsh (1991) showed that even under a variety of different matching strategies, matching designs are biased in favor of high achieving (gifted) classes due to differential regression to the mean. This means that when students from different ability groups (homogenous and high vs. heterogeneous or mixed; which are assumed to reflect true values) are matched on their Time 1 outcome measures and then compared with regard to their gains at Time 2, differential regression to the mean will lead to higher gain scores in the homogenous group, despite there being no differences in the true gain scores. This is because

the matched students from the heterogeneous group will show stronger regression to the mean, as the true mean of this group is lower on average than that of the high ability group, but they were matched to have the same level of outcome measure at Time 1.

In his seminal meta-analysis (including, correlational, experimental, and qualitative studies) on school compositional effects predicting achievement, Hattie (2002) concludes that ability tracking (i.e., grouping according to ability), although a widespread practice, has a close to zero effect. Hattie (2002) identifies instructional practices and teacher effects as overpowering any effects of school or class composition on achievement. He argued that any apparently small positive compositional effects are likely to be the result of uncontrolled variables (including differences in resources and curriculum, as well uncontrolled pre-treatment effects). He did, however, emphasize that the BFLPE was particularly robust. According to Hattie, a good teacher can achieve the same learning gains in a high or low ability classroom. Here the issue is on whether inequity (highly stratified school systems with large school-to-school variation in levels of achievement) results in excellence (higher levels of achievement) overall. Although the focus is not directly on composition effects per se, highly stratified school systems necessarily have many schools with high and low school-average achievement. Hence, the question becomes whether the existence of many schools with high school-average achievement results in higher test scores overall. However, research based on large cross-national data is increasingly showing that more segregated school systems result in lower – not higher—levels of achievement (Hanushek & Woesmann, 2005; Micklewright & Schnepf, 2007, OECD, 2013; Parker, et al., 2016; Willms, 2010). Rejecting the hypothesis of an efficiency/equity trade-off in academic performance, they found negative relations between performance and inequality that are robust and of statistical and practical significance. Although based on changes at the country level in large, cross-national studies, this research suggests that the policies that create schools with high and low levels of school-average achievement has negative effects on achievement overall.

The focus of our study is resolving the theoretical paradox through testing for a phantom effect that results in positively biased estimates of peer spillover, leading to the conclusion that attending a school with high-average achievement results in higher levels of achievement at the individual student level. Following from Harker & Tymms (2004), we consider the failure to control for measurement error and pre-existing differences as biases in the estimation of composition effects—the effects of school-average ability on ASC and achievement at the level of the individual student.

Measurement and sampling error. Measurement error with regard to compositional effects plays an interesting role. The unreliability of the individual level variable is naturally reduced in the aggregated version of the variable on the group level (as the aggregated variable reflects a group average). When both the individual and aggregated version of the variable, which are of course positively correlated, then predict a third variable, the prediction through the individual variable is negatively biased (opposite direction of the actual individual level estimate). The more reliable group level variable prediction, however, may be positively biased (in the direction of the actual within level estimate; e.g., Hauser, 1970; Televantou et al., 2015). This can be viewed as 'mopping up variance at the second level' (Harker & Tymms, 2004). Thereby, the biases of the group level effects are in the direction of the (positive) individual level effects.

It is assumed that a measure or scale only includes an unreliable subsample of items (analogous to people in the sampling error) that reflect the latent construct of interest, instead of the most likely infinite number of items that would be necessary to construe a perfectly reliable scale for assessing the construct (Televantou et al., 2015). Harker and Tymms (2004) found that the positive compositional effect of SES on achievement could be increased by deliberately adding unreliability to the SES measure on the individual level. Measurement error can be controlled, however, by utilizing structural equation modelling, which models multiple indicators of each (latent) factor (for an overview see Marsh et al., 2009; Pokropek,

2014). Indeed, research shows that so called doubly latent models (see Marsh et al., 2009), which correct for both measurement and sampling error, can reduce bias substantially (Pokropek, 2014; Televantou et al., 2015). Thus, Televantou and colleagues (2015) in a simulation showed that utilizing doubly latent models controlled appropriately for added measurement error. Moreover, in two studies based on different samples Televantou et al. (2015) demonstrated that their initially found positive peer spillover effect turned insignificant in one study and even significantly negative in the other after controlling for measurement error. Hence, in a study that does not control for measurement error there will be positive bias in the compositional effects on achievement and ASC, making the peer spillover effect more strongly positive than it should be and making the BFLPE seem to be less negative than it should be.

Additionally, measurement error in a multi-level model can arise from both sampling error and measurement error (Hutchinson, 2007, Televantou et al., 2015). Sampling error can affect group level effects by forming aggregates based on a (possibly unreliable) subsample of all individuals who are actually included in the group level units (Lüdtke et al., 2008; Marsh et al., 2009; Pokropek, 2014). However, Lüdkte et al., (2008) have introduced a multilevel latent approach, referred to as latent aggregation that takes into account sampling error when estimating group, and thus, compositional effects (see also Marsh et al, 2009). Put simply, latent aggregation can correct for sampling error (Televantou et al., 2015).

Correcting for Selection Effects by Controlling for Pre-existing Differences. A phantom effect can appear when researchers fail to include sufficient covariates in their models. From a methodological perspective, the effect of the group level predictor variable can be exaggerated due to pre-existing differences where variance in the outcome could be explained by selection effects (Hanushek, Kain, Markman, & Rivkin , 2003; Harker & Tymms, 2004; Hauser, 1970; Marks, 2015; Nash, 2003). This has been referred to as ecological or contextual fallacy (Hauser, 1970; Robinson, 2009; Slevin, 1958). Thus, Hauser

(1970) proposed that it is problematic to interpret residual differences between groups as social processes, as they are actually explained through the relationship of the residuals with relevant individual student-level predictors. Including these predictors will therefore reduce residual differences and possibly make them disappear altogether (leading to the term *phantom effect* – now you see it, now you do not). From a substantial perspective, this means that by omitting important variables, selection effects are neglected. In case of the peer spillover effect, the positive effect of school-level achievement on individual achievement could, for example, be explained through all students at a school coming from a high SES background which is usually associated with higher achievement levels (Jerrim, Parker, Chmielewski, & Anders, 2016).

Thus, in the case of investigating aggregated achievement it seems particularly important to include controls that are closely related with high achievement, such as strong measures of pre-treatment achievement collected prior to the introduction of the implicit intervention of attending a high-ability school, SES (or other available resources), gender, and ethnical heritage (Harker & Tymms, 2004; Marks, 2015; Strand, 2010; Jerrim et al.,2016).

Harker and Tymms (2004) found that after including prior achievement and ethnicity on the individual level (and cleaning the data of two outlier schools) the positive composite effect of SES on achievement disappeared. Further, Nash (2003) found that the positive compositional effect of SES in his data can be caused by omitted non-cognitive dispositions and variable family resources within social classes. Marks (2015) showed that the positive school composition effect of SES decreased after including prior ability on the individual level and school level, and even noted a small but statistically significant negative effect of school average ability (negative peer spillover effect) on later individual student ability when including prior achievement and SES in the model.

Overall these potential pitfalls of finding a phantom effect rather than a real compositional effect could apply for the BFLPE and the peer spillover effect alike. It is

important, however, to keep in mind that the nature of the phantom effect would work against the BFLPE in all cases (measurement error and lack of control for individual differences), as it biases estimates for group-level effects in the direction of the individual level effect (which is the opposite direction in case of the negative BFLPE). Put simply, a phantom effect would overestimate any positive peer spillover but underestimate a negative BFLPE. This highlights the apparent robustness of the BFLPE against such phantom effects, as there is strong evidence in its favor based on a wide range of research including studies that do not control for measurement or sampling error and additionally to not control for pre-existing differences. In contrast, the phantom effect is critical in the evaluation of peer spillover effects because control for the phantom effect will reduce the size of positive spillover effects, make them disappear altogether, or even shift the direction of the effect from positive to negative.

Attempts at Tackling the Theoretical paradox

So far our review of the literature points to strong evidence for the REM, strong evidence for a very robust BFLPE and apparently questionable evidence for the peer spillover effect. For investigating the apparent theoretical paradox of these models (see Figure 1), however, it is necessary to test all of the assumed relationships in an integrated model and applying a thorough methodology and design for avoiding any phantom effects. Indeed, as described earlier, the classic Marsh (1991) study found negative effects of school-average achievement on both ASC and academic achievement (based on both school grades and standardized achievement tests), and showed that the negative effect of school level ability on later GPA and individual test scores was mediated by ASC. Nevertheless, this early research predated subsequent methodological developments in appropriate handling of missing data and doubly-latent multilevel statistical models used in subsequent research.

More recently, researchers have addressed these issues, at least in part. Marsh and O'Mara (2010) controlled for measurement error and included a measure of SES in an 8-year longitudinal study. They found school-average achievement to negatively predict ASC and

GPA. Their study, however, did not incorporate multilevel modeling. Stäbler et al. (2016; also see Marsh et al., 2001) used multi-level modeling and found positive peer spillover, but this was based on manifest variables that did not control for measurement error. However, in an unpublished dissertation, Televantou (2014), controlled for measurement and sampling error in a study based on a sample of English students that included both the peer spillover effect (see also Televantou et al., 2015) and the BFLPE. In separate models for both compositional effects she found evidence that correcting for measurement error led to a more negative BFLPE and an insignificant or even negative peer spillover effect. When integrating both compositional effects in one model she found that the negative peer spillover could be explained in part by the BFLPE (Televantou, 2014).

The Present Study

The present investigation integrates into a single study tests of the REM, the BFLPE, and the peer spillover effect based on more sophisticated models to control for the phantom effect than have been used previously.

More specifically, we aim to test the validity of the REM and the two prominent school compositional effects based on school-level achievement, thereby attempting to resolve the theoretical paradox. Utilizing doubly latent multi-level modelling to correct for measurement and sampling error will correct for some unreliability and thus reduce the positive bias in compositional effects (e.g., Lüdtke, et al., 2008; Televantou et al., 2015). The present study will advance these findings by correcting for measurement error and including measures of prior mathematics achievement and important controls. We will then juxtapose these correctional measures by testing their effect on the BFLPE and the peer spillover effect.

Further, research has found a change of direction from positive to not significant or even significantly negative in the peer spillover effect, after correcting for measurement error (Televantou et al., 2015) and including omitted controls (Marks, 2015). This seems particularly important as it contradicts not only various research findings, but also policy and

parent decisions based on the assumption of a positive school composition effect on achievement. Televantou et al., (2015; also see Marsh, 1991) have provided some evidence that the peer spillover effect can in fact partly be explained by the BFLPE. Thus, it is assumed that ASC mediates the relationship of school average achievement on individual achievement. For clarity of presentation of our results however, the present study focusses on direct and total effects (the sum of all direct and indirect effects). For detailed results of the different indirect effects please see the Supplemental Material available online.

In detail we hypothesize the following:

Hypothesis 1(H1) – Baseline model:

In a baseline model including both school composition effects we predict:

(a) reciprocal effects on the individual level including positive cross-sectional as well as longitudinal cross-lagged relationships between ASC and academic achievement.

(b) negative compositional effects (direct, indirect via T2 ASC and achievement, and total) of school achievement on later individual ASC (BFLPE) and positive compositional effects (direct and total) of school achievement on later individual achievement (peer spillover effect) when the phantom effect is NOT controlled.

Hypothesis 2 (H2) – Controlling for the phantom effect:

In models correcting for unreliability and including strong controls for pre-existing differences we predict (for all direct and total effects):

(a) Correction of measurement error in student-level achievement will lead to a less positive (or even negative) peer spillover effects and more negative BFLPEs. The effects due to this correction will be small because the measure of achievement is highly reliable (see Method section).

(b) A decline or even negative effect of peer spillover effects and more

negative BFLPEs due to the inclusion of a measure of prior achievement as an additional indicator of the latent achievement variable.

(c) A decline or even negative effect of the peer spillover effects and more negative BFLPEs, after additionally including several achievement related controls (e.g., SES, gender, number of books at home).

(d) An even stronger decline or negative effect of the peer spillover effects and more negative BFLPEs when combining (a)+(b)+(c). This decline of the peer spillover effects and more negative BFLPEs should be larger when all the controls are used in a single model than models considering each of them separately.

Method

Participants

Participants were 21,260 students (a nationally representative sample of US kindergartners) who participated in the Early Childhood Longitudinal Study Kindergarten Class of 1998-99 (ECLS-K; Tourangeau, Nord, Lê, Sorongon, and Najarian, 2009; for details see Supplemental Material available online). We consider longitudinal data from spring-kindergarten (TK), spring-first grade (T1), spring-third grade (T2), and spring-fifth grade (T3). Consistent with previous research (Marsh & Hau, 2003; Seaton, Marsh, & Craven, 2010) we included only schools with at least 10 students to ensure a reliable estimate of school-average achievement. The resulting final sample consisted of 14,985 students from 853 schools. The average sample size per school was 15 students; 57% of the students were white (non-Hispanic), 14% were black, 17% were Hispanic, 6% were Asian, and 6% were from other ethnic backgrounds. Further, 51% of the sample was female. Average age was 7.24 years (SD = 0.35) at spring-first grade assessment (T1). Based on a categorical variable of SES (including five categories) 18 % of students were in the first, 19% in the second, 20%

in the third, 21 % in the fourth, and 22% in the fifth quintile.

Measures

Table S1 in the Supplemental Material available online presents the latent correlations of all factors based on a CFA.

Mathematics achievement. Mathematics achievement was assessed through a standardized test on every measurement occasion (for details see Supplemental Material available online). The internal consistencies of these scales are reported to be very high with .93 at TK (this is the early Kindergarten achievement measure), .94 at T1, .95 at T2, and .95 at T3, i.e., there is very little measurement error to control for. Intraclass correlation coefficients (ICC-1s) revealed a substantive amount of school-to-school variations in achievement; .21 at T1, .22 at T2, and .22 at T3.

Self-concept. Academic self-concept in mathematics (ASC) was assessed with three items of the Self-Description Questionnaire-I (SDQ-I; Marsh, 1990) assessing perceived competence in math at T2 and T3; self-concept was not measured at T1. All items (e.g., "I am good at math") of this scale were measured with a 4-point Likert scale (1= "not at all true" to 4 = "very true"). The internal consistencies of this scale were high with .81 at T2, and .86 at T3. In contrast to the achievement scores, ICCs for ASCs were small; overall ranging from .02 to .04.

Covariates. As covariates we included gender, age, ethnic heritage (dummy coded with white non-Hispanic as reference category, then Black, Hispanic, Asian/ Pacific Islander, and other as remaining categories), and several variables reported by the parents: how many books the student had at home (at T1), how often the student reads outside of school (at T1), and SES (see Supplemental Material available online for further discussion).

Statistical analysis

Generally, given the known sensitivity of the chi-square test to sample size, to minor deviations from multivariate normality, and to minor misspecifications, applied SEM research

focuses on indices that are relatively sample-size independent (Hu & Bentler, 1999; Marsh, Hau, & Wen, 2004), such as the Root Mean Square Error of Approximation (RMSEA), the Tucker-Lewis Index (TLI), and the Comparative Fit Index (CFI). Population values of TLI and CFI vary along a 0-to-1 continuum, in which values greater than .90 and .95 typically reflect acceptable and excellent fits to the data, respectively. Values smaller than .08 and .06 for the RMSEA support acceptable and good model fits, respectively. As is typical in large longitudinal field studies, a substantial portion of the sample had missing data for at least one of the measurement waves, due primarily to absence or students changing schools. The majority of students completed all three waves of data (64.6%), while 19.8% and 15.5% of students only completed two or one wave of data, respectively. Missing data was handled using the full information maximum likelihood (FIML) approach (Enders, 2010).

All models were analyzed as random intercept multi-level models including two levels: Level 1(L1) = individual student, Level 2 (L2) = school. This modeling approach takes into account the nested (non-independent) structure of the data (students nested within schools), thus rendering standard errors that are corrected for this nesting.

The simplest equation for such random intercept models is generally depicted as:

$$Y_{ij} = \beta_{00} + \beta_{10}(X_{ij} - \bar{X}_{.j}) + \beta_{01}\bar{X}_{.j} + \delta_{0j} + \varepsilon_{ij} \quad (1)$$

Where Y_{ij} is the outcome for student i in school j, β_{00} is the grand-mean intercept, β_{10} is the within-group regression coefficient describing the relationship between X_{ij} and Y_{ij} within each group, β_{01} is the between-group regression coefficient describing the relationship of school means. The parameters δ_{0j} and ε_{ij} represent the standard normally distributed uncorrelated residuals at the school and student level, respectively.

In our models for correcting for sampling error, which can result when only a limited number of students (in case of the present study approx.. 25 per school) are sampled from each school, we applied latent aggregation in all our models (see Lüdtke et al., 2008; Marsh et

al., 2009). Moreover, in all our models we estimated latent variables, thereby correcting for measurement error in our ASC measure (see Supplemental Material for details). In some subsequent models we additionally corrected for measurement error in our achievement measure (see below). Models that simultaneously control for measurement and sampling error can be referred to as doubly latent based on Marsh et al.'s (2009) 2 x 2 taxonomy of multi-level models. Thus, these models consist of several manifest indicators (l = 1, . . . , L; or k = 1, . . . , K) per latent factor rather than a composite scale score (usually an average of all items), correcting for measurement error. Additionally the L2 variables are not simple aggregates of the L1 variables, but latent variables corrected for sampling error. The equation for such models can be depicted as follows (see Marsh et al., 2009 for details):

For the indicators of the outcome variable (measurement model):

$$\mathrm{Y_{lij}} = \mu_{\mathrm{ly}} + \lambda_{\mathrm{ly},W} U_{yij} + \mathrm{R_{lyij}} + \lambda_{\mathrm{ly},B} U_{yj} + \mathrm{R_{lyj}}; \qquad l = 1, \ldots \ldots, L \tag{2}$$

For the indicators of the predictor variable (measurement model):

$$\mathrm{X_{lij}} = \mu_{\mathrm{kx}} + \lambda_{\mathrm{kx},W} U_{xij} + \mathrm{R_{kxij}} + \lambda_{\mathrm{kx},B} U_{xj} + \mathrm{R_{kxj}}; \qquad k = 1, \ldots \ldots, K \tag{3}$$

In these equations (where y can be replaced for x depending on outcome or predictor variable) $\lambda_{\mathrm{ly},W}$ are the within factor loadings, U_{yij} reflects the according true score on L1, and $\mathrm{R_{lyij}}$ is the residual at L1. Similarly $\lambda_{\mathrm{ly},B}$reflects the between factor loadings, U_{yj}. The L2 true score and $\mathrm{R_{lyj}}$ the L2 residual. The structural model is then similar to the initial simple random intercept model (see Equation[1]):

$$U_{yij} = \beta_{00} + \beta_{10} U_{xij} + \beta_{01} U_{xj} + \delta_{0j} + \varepsilon_{ij} \tag{4}$$

In the present study we used Mplus 7.4 (Muthén & Muthén, 1998–2010) and its model constraint option to calculate all compositional effects and their standard errors. For estimating effect sizes we used the most conservative effect size measure recommended by Parker, Marsh, Lüdtke, & Trautwein (2013) based on the total L1 variance. We investigated Lag 1 compositional effects (T1-T2) as well as Lag 2 compositional effects (T1-T3). We used

the same approach as Marsh et al. (2016) where in our case, Lag 2 compositional effects are the effects of school composition after controlling for Lag 1 compositional effects as well as the effects of L1 variables from the earlier waves. Thus, significant compositional effects at Lag 2, of the same direction as the Lag 1 effect, would indicate "sleeper effects" (new effects, in addition to the effects already observed, these could be either positive or negative; see also Marsh et al., 2016). Nonsignificant compositional effects at Lag 2 would indicate that Lag 1 compositional effects were maintained, and significant compositional effects at Lag 2 of the opposite direction of the Lag 1 effect would indicate that Lag 1 compositional effects were not fully maintained.

For investigating possible bias due to the phantom effect we present an a priori series of sequential models (For the Mplus Syntax of our models please see Supplemental Material available online):

1. We modeled our baseline model which included mathematics achievement at T1, T2, and T3 as well as MSC at T2 and T3. L2 achievement at T1 predicted L1 self-concept (BFLPE) and L1 achievement (Peer spillover effect). In addition this model included reciprocal effects at L1 (see Figure 2).
2. In Model 2 we added controls for unreliability and pre-existing differences to Model 1 by: (a) correcting for measurement error in the achievement variables by modeling a latent achievement variable with one indicator and a factor loading constrained to one (see Supplemental Material available online, Model 2a); (b) including earlier achievement in Kindergarten (TK) as a second manifest indicator for the latent baseline achievement variable[2] (Model 2b); and (c) including all of our covariates as predictors on L1 in order to control for pre-existing differences (Model 2c).
3. We then set up a combination Model 3 where we integrated the models 2a-2c. In Model 3 we simultaneously corrected for measurement error, included

kindergarten achievement as an additional indicator for prior achievement, and included the other achievement related controls (Model 2a+b+c).

Results

Baseline Model (H1)

In our baseline model we modeled all four school compositional effects simultaneously (see Figure 2). Thus, we included mathematics achievement at T1, T2, and T3 as well as ASC in Mathematics at T2 and T3. The model fitted the data well (CFI > .99, TLI = .99, RMSEA = .02; for an overview of model fit for all models see Table 1, for the Mplus Syntax of all models please see Supplemental Material available online).

Reciprocal effects (H1a). For accurately representing the reciprocal relationship of academic achievement and self-concept, we included reciprocal effects in our baseline model. Both the paths from mathematics achievement to ASC and the path from ASC to math achievement were confirmed. Thus, math achievement at T2 and T3 was positively related to ASC at both T2 (B =.10, p<.001) and T3 (B = .06, p<.001), respectively. Additionally, achievement at T1 positively predicted achievement at T2 (B = .75, p<.001) and T3 (B =.15, p < .001) as well as ASC at T2 (B = .25, p<.001) and T3 (B = .06, p<.01). The stability (test-retest T2-T3) paths of achievement and self-concept were also both positive with B=.74, p < .001 and B = .40, p < .001, respectively. Most importantly, however, are the cross effects: Achievement at T2 positively predicted ASC at T3 (B = .29, p < .001) while ASC at T2 significantly and positively predicted achievement at T3 (B = .02, p < .01). These results in support of the reciprocal effects model remained stable across all models.

Compositional effects (H1b). The model included four longitudinal direct compositional effects (from T1 to T2 and from T1 to T3): Two relating school-average achievement to ASC (BFLPEs) and two relating school-average achievement to individual student achievement (peer spillover effect). The Lag 1 BFLPE (T1-T2) was significantly

negative, while the direct Lag 2 BFLPE was also negative but not significant. Overall this pattern indicates that the BFLPE was maintained over time.

The Lag 1 peer spillover effect was significantly positive. The direct Lag 2 peer spillover effect was significantly negative. The total Lag 2 effect was negative. As this Lag 2 effect was controlled for the Lag 1 peer spillover effect, this effect of opposite sign than the Lag 1 effect indicated that the Lag 1 effect on achievement was not maintained (see Table 2).

Models Correcting for Phantom Effects (H2)

In a next step we tested the three models correcting for unreliability in the achievement measures and including pre-existing differences as controls.

Correcting for measurement error (H2a). In the first of these models (Model 2a) we corrected all achievement scores for measurement error (see Supplemental Material available online), while the setup of the model was identical to the baseline model (Model 1). Because only one indicator with a fixed loading was used for each achievement measure, this addition necessarily had no effect on model fit. The pattern of direct, indirect, and total effects was also very similar to that of the baseline model. Regarding the compositional effects, the Lag 1 BFLPE (T1-T2) was significantly negative. The Lag 2 direct BFLPE remained non-significant. The Lag 1 peer spillover effect was significantly positive. The direct Lag 2 peer spillover effect was negative, but not significant, while there was a positive total effect, indicating that the Lag 1 effect on achievement was maintained (see Table 2).

Adding Kindergarten achievement as an indicator (H2b). In the second of these models (Model 2b) we added prior achievement at TK (Kindergarten) as an additional manifest indicator for a latent achievement variable at T1, while the rest of the model was again identical to the baseline model (Model 1). This addition slightly improved model fit CFI > .99, TLI = .99, RMSEA = .02). In this model the Lag 1 BFLPE (T1-T2) was again significantly negative and increasingly so while the direct Lag 2 BFLPE remained non-significant. However, the Lag 2 BFLPE was also slightly more negative than in Model 2a.

Further, results showed a negative and significant total effect, indicating an additional negative effect of aggregated achievement over and beyond the Lag 1 BFLPE (see Table 2).

The Lag 1 peer spillover effect was no longer significant and close to zero (see Table 2). The direct Lag 2 peer spillover effect remained negative and was significant as did the total effect. This indicated that there was no significant effect of school-average achievement on individual achievement at Lag 1, but a significantly negative effect total effect at Lag 2 (i.e, a negative sleeper effect).

Including controls (H2c). In a third model (H2c) we again used the identical set up as the baseline model, but predicted all L1 variables by a set of controls for pre-existing differences (i.e., gender, age, how often the student reads outside of school, how many books the student has at home, race, and SES; Model 3)[3]. The fit of this model was good CFI = .99, TLI = .98, RMSEA = .02. Including these controls lead to an even more negative and significant Lag 1 BFLPE (T1-T2) and direct Lag 2 BFLPE. Thus, the total effect was also significant and even more negative than before, indicating an ongoing negative effect (see Table 2).

The Lag 1 peer spillover effect was now also negative, although not significant. The direct Lag 2 peer spillover effect remained negative and significant as did the total effect, indicating a negative sleeper effect in addition to the non-significant negative Lag 1 effect on achievement (see Table 2).

SES showed significant positive effects on achievement. The other controls for pre-existing differences included in this model (gender, ethnic heritage, books at home, and reading outside of school) revealed relatively small, mostly non-significant effects (see Supplemental Materials available online and Table 3).

Combination Model (H3)

In the last model we combined all three measures to correct for the phantom effect. Model 3 showed good fit (Table 1). Both the Lag 1 BFLPE and the Lag 1 spillover effect

were negative and significant. Both direct Lag 2 compositional effects were negative but insignificant. All total effects were, however, significant and again more negative than in any prior model (see Table 2). Hence, both Lag 1 compositional effects were now significantly negative. Again the BFLPE Lag 2 showed an additional, even stronger total negative effect over and beyond the Lag 1 effects. The Lag 2 peer spillover effects remained negative, but not significant indicating maintenance of the Lag 1 effect.

Overall our results showed consistently negative and significant BFLPEs for Lag 1 and mostly negative and significant Lag 2 BFLPEs total effects. Correcting for phantom effects led to more negative BFLPE effects. The Lag 1 peer spillover effect decreased and even became negative depending on the various controls for the phantom effect. The Lag 2 total compositional effect on achievement remained negative for all models except one. Put simply, consistent with a priori predictions, control for the phantom effect (correcting for measurement error and controlling for pre-existing differences) led to increasingly negative total BFLPEs and less positive total spillover effects that became negative when all controls were included. These results not only resolve the theoretical paradox, but have important implications for the organization of schooling.

Discussion

The major aim of the present study was to investigate an apparent theoretical paradox in the juxtaposition of the REM, the BFLPE and the peer spillover effect—particularly the seemingly inconsistent effects of school-average achievement on ASC and achievement at the individual student level. Based on our literature review the key to explaining this paradox lies in unmasking phantom effects masquerading as positive peer spillover effects and leading to understating the size of the negative BFLPE.

We began by replicating support for the REM, the first link in our theoretical paradox. Consistent with Hypothesis 1a (H1a) and previous REM research (Marsh & Craven, 2006; Seaton, Marsh, et al., 2015; Valentine et al., 2004), we found that at the individual student

level ASC and achievement are reciprocally related and mutually reinforcing such that positive changes in one leads to positive changes in the other. This would seem to suggest that conditions leading to changes in either achievement or ASC would have similar effects on the other. Herein lies the seeds of our theoretical paradox. Indeed, based on traditional approaches to compositional effects, we found apparent support for the theoretical paradox—the inconsistent effects of school-average achievement on ASC and achievement. In support of our a priori predictions (H1b) when the phantom effect was not controlled, we found school-average achievement had short-term positive effects on individual student achievement but negative effects on ASC even though achievement and ASC were reciprocally and positively related. Total long-term effects of both however, were negative throughout all models except one.

In the next link in resolving the paradox and again consistent with a priori hypotheses, correcting for measurement error led to slightly more negative BFLPE effects, and slightly less positive peer spillover effects (H2a; also see Harker & Tymms, 2004; Pokropek, 2014; Televantou et al., 2015). Nevertheless, these effects of controlling for measurement error were small because the achievement test scores already had high reliability (Tourangeau et al., 2009).

The final links to understanding the paradox consisted of the addition of prior achievement (H2b) and covariates (H2c).When we controlled for these sources of the phantom effect the BFLPE became even more negative while the short-term (direct) peer spillover effect went from positive to nearly zero, in line with Televantou et al.'s (2015) findings (H2c; see also: Harker & Tymms, 2004; Hauser, 1970; Marks, 2015; Nash, 2003).

Not surprisingly, the combination of controlling for all sources of the phantom effect in the same model led to an even stronger decrease of the peer spillover effect, and also to an even more negative BFLPE (H2d). This shows the importance of both aspects - unreliability and the omission of covariates – responsible for phantom effects (Harker & Tymms, 2004).

Although control for pre-existing differences was more important than control for measurement error in our study, this is largely due to the highly reliable measures (and additionally correcting for sampling error in all models) of achievement in our study. Indeed, control for measurement error is likely to have substantially larger effects in studies where the achievement test scores are not already highly reliable (Televantou et al., 2015; also see Stäbler et al., 2016). Likewise the effect of controlling pre-existing differences will depend in part of the magnitude of these differences and the collection of variables to represent them. Ultimately, without random assignment, the positive bias of pre-existing differences in favor of students attending schools with high school-average achievements can never be completely eliminated. For this reason, we suspect that even the close to zero effects of school-average achievement on individual student achievement in the present study are positively biased, and would become more negative if we had a better set of covariates to control more of the pre-existing differences that favor students from schools with high school-average achievement. This highlights the importance of well-designed and thoroughly conducted research that simultaneously controls for measurement error as well as controlling the inevitable pre-existing differences in favor of students in high-achieving schools (Goldring, 1990).

Resolving the theoretical paradox

Taken together, the peer spillover effect seems to be explained through measurement error, and relevant individual student-level predictors, rather than by a true compositional effect of achievement (Hauser, 1970). The BFLPE, however, becomes even more negative after controlling for measurement error, prior achievement, and the covariates. Thus, the BFLPE remains a robust compositional effect, while we found strong evidence against the validity of a positive peer spillover effect.

Moreover, our findings resolve the theoretical paradox juxtaposing the REM (positive reciprocal effects of individual achievement and ASC over time), the negative BFLPE (the negative effects of group level achievement on individual ASC), and apparently positive peer

spillover effects (the positive effects of group level achievement on subsequent individual level achievement). Introducing stronger statistical models than considered in previous research shows the actual effects of school-average achievement on individual self-concept and achievement are in harmony with what would be expected on the basis of the BFLPE and REM models. The present investigation shows that with appropriate statistical controls, the effect is even slightly below zero.

Strengths, Limitations, and Future Directions

The present investigation has a number of important strengths, including the application of theory driven longitudinal models, the utilization of strong (doubly latent) multi-level analysis (compositional effects), and use of a large, nationally representative sample. Nevertheless, some limitations need to be addressed.

Although including numerous achievement related covariates, there are inevitably other omitted covariates that could control for additional pre-existing differences not considered in our study (Hanushek et al , 2003; Harker & Tymms, 2004; Hauser, 1970) and should be included in future analysis. Consistent with our rationale, these would be expected to make both the BFLPE and peer spillover effect even more negative.

Further, while most regression coefficients in our models were statistically significant, some showed rather small effect sizes. Our main point, however, is not the effect sizes, but that the direction of peer spillover effect is not positive, which is most important from a policy perspective. The generic nature of the achievement test that was central to the present investigation is both a strength and a weakness. On the one hand, it was specifically designed to be appropriate across a wide range of ability levels rather than a particular curriculum. On the other hand, these standardized achievement tests were "low-stakes" tests in that they have no consequences for individual students in terms of school marks, feedback to parents, or progression in a student’s educational career; students have no way to prepare for them and do not receive feedback on their results. Thus, it is important to test the generalizability of our

results in studies based on a variety of different measures of achievement (high and low stakes tests); standardized test scores; tests designed to be curriculum-specific and curriculum-generic (see related discussion; e.g., Marsh & Seaton, 2015).

Practical and Policy implications

The results of the present study call into question previous studies and associated reviews showing positive effects of school-average achievement, tracking, and/or academically selective schools on individual achievement (Harker & Tymms, 2004; Nash, 2013). Similarly, meta-analyses based on such research experience this issue (see earlier discussion of Goldring, 1990) in that they are largely based on studies that fail to control for a phantom effect. Thus, subsequent systematic reviews and meta-analyses need to re-evaluate all previous research in relation to appropriate controls for any phantom effects.

Furthermore, our results question how meaningful or maybe even harmful educational policies may be, as these are largely based on biased studies that have wrongfully found a positive peer spillover effect (Nash, 2013). Although promoting the achievement of all students should be the primary aim of formal schooling, placing students in high-achievement schools can have negative and detrimental side effects (when excluding other important schools related factors – see below for more details), as demonstrated here.

Moreover for individual parents, this means sending your children to high achievement schools likely does not lead to high returns on investment, at least with regard to achievement and ASC. On the contrary, being placed in a high achieving group of students will have a negative impact on an individual student's ASC and no positive or even slightly negative effect on their achievement as well. As individual achievement and individual self-concept are reciprocally related, a negative loss spiral of a student's motivation and achievement could occur. However, such recommendations need to be viewed taking into account that we estimate average effects. As a result, policy implications refer to a macro level, while transferability of recommendations to individual children is limited. Furthermore,

our findings are consistent with Hattie‘s work (2002, 2009) mentioned above. In his meta-analyses he showed that there is little or no evidence for the benefits of ability stratification even for achievement measures. Instead, Hattie (2002, 2009) proposes that the positive effects reported for gifted programs are not due to ability tracking itself, but to improved curriculum and quality of education (Marsh, Kuyper, Morin, Parker, & Seaton, 2014). Expanding upon this theme, here we show that many of the studies included in these meta-analysis were systematically biased in the direction of positive peer spill-over effects and that if controls for a phantom effect had been introduced the overall effect sizes would likely to have been negative rather than positive. However, we agree with Hattie (2009; see also Marsh, Kuyper et al., 2014), who suggested that most of the features of educational programs for talented students reflect those educational practices that would provide a similar benefit to average-ability students and students in homogeneous classes. Research providing alternative perspectives based on strong theory and state of the art methodology, i.e. unmasking a phantom effect, seems especially important in this regard, as so many teachers, parents, and policy makers uncritically assume that a stratified school environment will automatically benefit the attending students. Thus, our results add to the literature that promotes educational equality over ability stratification for fostering students’ achievement and ASC (see Parker et al, 2016 for an overview).

It is important to emphasize that the focus of the present investigation is on school composition effects rather than the many other differences between schools with high-achieving students and those with mixed-ability and lower-achieving students. Indeed, many of these extra-compositional differences are likely to favor high-achieving schools, such as elite private schools (e.g., teaching resources, school/classroom facilities, per-pupil expenditure, extracurricular enrichment activities, teacher salaries, networks, experience, levels of training etc.), and even local social and economic factors (Jerrim, Parker, Chmielewski, & Anders, 2016). However, based on our results we assume that additionally

controlling for these (pre-existing) differences will most likely also result in more negative effects.

Distinguishing pure compositional effects associated with the ability level of students within a school from the extra-compositional effects that are likely to be confounded with school-average achievement, is beyond the scope of the present investigation. However, it is an important direction for future research, particularly in relation to the organization of schools. More specifically, if effects associated with high-achieving schools are the net effects of even more negative compositional effects than we have found here and possible positive effects of extra-compositional variables, this would argue against inequitable stratification in school systems which results in high numbers of high-achieving schools (and, by necessity, also many low-achieving schools). Although support for this supposition is beyond the scope of the present investigation, it is consistent with the growing support for the finding that more equitable, less stratified school systems result in better overall achievement (OECD, 2013; Parker et al., 2016; also see Salchegger, 2016). An exciting direction of future research is to bring together these two different strands of research arguing for more equitable school systems and against the stratification of schools based on student achievement (or variables highly related to achievement), including academically selective schools for high-achieving students. Put simply, less stratified school systems may raise all boats.

References

Becker, M., & Neumann, M. (2016). Context-related changes in academic self-concept development: On the long-term persistence of big-fish-little-pond effects. *Learning and Instruction, 45*, 31–39. https://doi.org/10.1016/j.learninstruc.2016.06.003

Calsyn, R. J., & Kenny, D. A. (1977). Self-concept of ability and perceived evaluation of others: Cause or effect of academic achievement?. *Journal of Educational Psychology, 69*, 136–145. https://doi.org/10.1037/0022-0663.69.2.136

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

Craven, R. G., & Marsh, H. W. (2000). Gifted, streamed and mixed-ability programs for gifted students: Impact on self-concept, motivation, and achievement. *Australian Journal of Education, 44*, 51–75. https://doi.org/10.1177/000494410004400106

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. https://doi.org/10.1080/0141192032000137330

Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford Press.

Goetz, T., Pekrun, R., Zirngibl, A., Jullien, S., Kleine, M., vom Hofe, R., & Blum, W. (2004). Leistung und emotionales Erleben im Fach Mathematik: Längsschnittliche MehrebenenanalysenAcademic Achievement and Emotions in Mathematics: A Longitudinal Multilevel Analysis Perspective. *Zeitschrift für Pädagogische*

Psychologie, 18(3/4), 201–212. https://doi.org/10.1024/1010-0652.18.34.201

Goldring, E. B. (1990). Assessing the status of information on classroom organizational frameworks for gifted students. *The Journal of Educational Research, 83*, 313–327. https://doi.org/10.1080/00220671.1990.10885977

Hanushek, E. A., Kain, J. F., Markman, J. M., & Rivkin, S. G. (2003). Does peer ability affect student achievement? *Journal of Applied Econometrics, 18*, 527–544. https://doi.org/10.1002/jae.741

Hanushek, E.A. & Wößmann, L. (2005). Does educational tracking affect performance and inequity? Differences-in-differences evidence across countries. *The Economics Journal, 116*, 63–76. https://doi.org/10.1111/j.1468-0297.2006.01076.x

Harker, R., & Tymms, P. (2004). The effect of student composition on school outcomes. *School Effectiveness and Improvement, 15*(2), 177–199. https://doi.org/10.1076/sesi.15.2.177.30432

Hattie, J. A. (2002). Classroom composition and peer effects. *International Journal of Educational Research, 37*, 449-481. https://doi.org/10.1016/S0883-0355(03)00015-6

Hattie, J. A. C. (2009). Visible learning: A synthesis of over 800 meta-analyses relating to achievement. London, UK: Routledge.

Hauser, R. M. (1970). Context and consex: a cautionary tale. *American Journal of Sociology,* 645–664. https://doi.org/10.1086/224894

Hu, L. T., & 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, 1–55. https://doi.org/10.1080/10705519909540118

Huang, C. (2011). Self-concept and academic achievement: A meta-analysis of longitudinal relations. *Journal of School Psychology*, *49*, 505–528. https://doi.org/10.1016/j.jsp.2011.07.001

Hutchison, D. (2007). When is a compositional effect not a compositional effect? *Quality & Quantity*, 41(2), 219–232. https://doi.org/10.1007/s11135-007-9094-2

Jerrim, J., Parker, P. D., Chmielewski, A. K., & Anders, J. (2016). Private schooling, educational transitions, and early labour market outcomes: Evidence from three Anglophone countries. *European Sociological Review, 32*, 280–294. https://doi.org/10.1093/esr/jcv098

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: a new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, *13*(3), 203–229. https://doi.org/10.1037/a0012869

Marks, G. N. (2010). What aspects of schooling are important? School effects on tertiary entrance performance. *School Effectiveness and School Improvement*, *21*, 267–287. https://doi.org/10.1080/09243451003694364

Marks, G. N. (2015). Are school-SES effects statistical artefacts? Evidence from longitudinal population data. *Oxford Review of Education, 41*, 122–144. https://doi.org/10.1080/03054985.2015.1006613

Marsh, H. W. (1990). *Self Description Questionnaire III: SDQ III Manual*. University of Western Sydney, Macarthur.

Marsh, H. W. (1991). The failure of high ability high schools to deliver academic benefits: The importance of academic self-concept and educational aspirations. *American*

Educational Research Journal, 28, 445–480. https://doi.org/10.3102/00028312028002445

Marsh, H. W. (2007). *Self-concept theory, measurement and research into practice: The role of self-concept in Educational Psychology*. Leicester, UK: British Psychological Society.

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*, 133–2163. 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*(5), 364–376. https://doi.org/10.1037/0003-066X.58.5.364

Marsh, H. W., Hau, K. T., & 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), 320-341. https://doi.org/10.1207/s15328007sem1103_2

Marsh, H. W., Köller, O., & 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), 321–350. https://doi.org/10.3102/00028312038002321

Marsh, H. W., Kuyper, H., Morin, A. J., 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., 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*, 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*, 764–802.
https://doi.org/10.1080/00273170903333665

Marsh, H. W., Nagengast, B., Fletcher, J., & Televantou, I. (2011). Assessing educational effectiveness: Policy implications from diverse areas of research. *Fiscal Studies, 32*, 279–295.
https://doi.org/10.1111/j.1475-5890.2011.00137.x

Marsh, H. W., & O'Mara, A. (2008). Reciprocal effects between academic self-concept, self-esteem, achievement, and attainment over seven adolescent years: Unidimensional and multidimensional perspectives of self-concept. *Personality and Social Psychology Bulletin, 34*(4), 542–552.
https://doi.org/10.1177/0146167207312313

Marsh, H. W., & O'Mara, A. J. (2010). Long-term total negative effects of school-average ability on diverse educational outcomes. *Zeitschrift für Pädagogische Psychologie. 24*, 51–72.
https://doi.org/10.1024/1010-0652/a000004

Marsh, H. W., Pekrun, R., Parker, P. D., Murayama, K., Guo, J., Dicke, T., & Lichtenfeld, S. (2016). Long-Term Positive Effects of Repeating a Year in School: Six-Year Longitudinal Study of Self-Beliefs, Anxiety, Social Relations, School Grades, and Test Scores. *Journal of Educational Psychology.* Advance online publication. https://doi.org/10.1037/edu0000144

Marsh, H. W., Trautwein, U., Lüdtke, O., & Köller, O. (2008). Social comparison and big-fish-little-pond effects on self-concept and other self-belief constructs: Role of generalized and specific others. *Journal of Educational Psychology, 100*, 510–524. https://doi.org/10.1037/0022-0663.100.3.510

Marsh, H. W., & Shavelson, R. (1985). Self-concept: Its multifaceted, hierarchical structure. *Educational Psychologist, 20*, 107–123. https://doi.org/10.1207/s15326985ep2003_1

Marsh, H. W., & Seaton, M. (2015). The Big-Fish–Little-Pond Effect, Competence Self-perceptions, and Relativity: Substantive Advances and Methodological Innovation. *Advances in Motivation Science*, 2, 127–184. https://doi.org/10.1016/bs.adms.2015.05.002

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(3)*, 319–350. https://doi.org/10.1007/s10648-008-9075-6

McDonald, R. P., & Marsh, H. W. (1990). Choosing a multivariate model: Noncentrality and goodness-of-fit. *Psychological Bulletin, 107,* 247–255. https://doi.org/10.1037/0033-2909.107.2.247

Mickelwright, J. & Schepf, S. (2007). Inequalities in industrialised countries. In S.P. Jenkins & J. Micklewright (Eds). *Inequality and Poverty Re-examined* (pp. 129-145). Oxford, UK: Oxford University Press.

Muthén, L. K., & Muthén, B. O. (1998–2010). *Mplus users guide*. Los Angeles, CA: Muthén & Muthén.

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*, 1033-1053. https://doi.org/10.1037/a0027697

Nash, R. (2003). *Is the school composition effect real? A discussion with evidence from the UK PISA data.* School Effectiveness and Improvement, 14(4), 441–457. https://doi.org/10.1076/sesi.14.4.441.17153

OECD (2013), "Equity in Outcomes", in PISA 2012 *Results: Excellence through Equity (Volume II): Giving Every Student the Chance to Succeed*, OECD Publishing, Paris. https://doi.org/10.1787/9789264201132-7-en

Parker, P. D., Jerrim, J., Schoon, I., & Marsh, H. W. (2016). A Multination Study of Socioeconomic Inequality in Expectations for Progression to Higher Education. *American Educational Research Journal, 53*, 6–32. https://doi.org/10.3102/0002831215621786

Parker, P. D., Marsh, H. W., Lüdtke, O., & Trautwein, U. (2013). Differential school contextual effects for math and English: Integrating the big-fish-little-pond effect and the internal/external frame of reference. *Learning and Instruction*, *23*, 78–89. https://doi.org/10.1016/j.learninstruc.2012.07.001

Pokropek, A. (2015). Phantom Effects in Multilevel Compositional Analysis Problems and Solutions. *Sociological Methods & Research*, *44,* 677–705. https://doi.org/10.1177/0049124114553801

Robinson, W. S. (2009). Ecological correlations and the behavior of individuals. International Journal of Epidemiology, 38(2), 337–341.

Salchegger, S. (2016). Selective school systems and academic self-concept: How explicit and implicit school-level tracking relate to the big-fish-–little-pond effect across cultures. *Journal of Educational Psychology*, 108, 405–423. https://doi.org/10.1037/edu0000063

Seaton, M., Marsh, H. W., & Craven, R. G. (2010). Big-Fish-Little-Pond Effect Generalizability and Moderation—Two Sides of the Same Coin. *American Educational Research Journal, 47*, 390–433. https://doi.org/10.3102/0002831209350493

Seaton, M., Marsh, H. W., Parker, P. D., Craven, R. G., & Yeung, A. S. (2015). The Reciprocal Effects Model Revisited Extending Its Reach to Gifted Students Attending Academically Selective Schools. *Gifted Child Quarterly*, 59(3), 143–156. https://doi.org/10.1177/0016986215583870

Selvin, H. C. (1958). Durkheim's suicide and problems of empirical research. American Journal of Sociology, 63(6), 607–619.

Stäbler, F., Dumont, H., Becker, M., & Baumert, J. (2016). 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*. Advance online publication. https://doi.org/10.1037/edu0000135

Strand, S. (2010). Do some schools narrow the gap? Differential school effectiveness by ethnicity, gender, poverty, and prior achievement. *School Effectiveness and School Improvement, 21*, 289–314. https://doi.org/10.1080/09243451003732651

Televantou, I. (2014). *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* (Doctoral dissertation, University of Oxford).

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*, 75–101. https://doi.org/10.1080/09243453.2013.871302

Thrupp, M., Lauder, H., & Robinson, T. (2002). School composition and peer effects. *International Journal of Educational Research, 37*, 483–504. https://doi.org/10.1016/S0883-0355(03)00016-8

Tourangeau, K., Nord, C., Lê, T., Sorongon, A. G., and Najarian, M. (2009). *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Combined User's Manual for the ECLS-K Eighth-Grade and K–8 Full Sample Data Files and Electronic Codebooks (NCES 2009–004).* National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Valentine, J. C., DuBois, D. L., & Cooper, H. (2004). The relation between self-beliefs and academic achievement: A meta-analytic review. *Educational Psychologist, 39*, 111–133. https://doi.org/10.1207/s15326985ep3902_3

Willms, J. D. (1985). The balance thesis: contextual effects of ability on pupils' O‐grade examination results. *Oxford Review of Education, 11*, 33–41. https://doi.org/10.1080/0305498850110103

Zell, E., & Alicke, M. D. (2009). Contextual neglect, self-evaluation, and the frog-pond effect. *Journal of Personality and Social Psychology, 97*, 467–482. https://doi.org/10.1037/a0015453

Zell, E., & Alicke, M. D. (2010). Personality and Social Psychology. *Personality and Social Psychology Review, 14*(4), 368–384. https://doi.org/10.1177/1088868310366144

Footnotes

[1] This effect is also referred to as contextual effect (e.g., Willms, 1985). However, Harker & Tymms (2004) argue that strictly contextual effects include other variables (such as governance structures, grade levels, and size) than those that make up a schools composition.

[2] We chose including TK achievement as a manifest indicator rather than as a covariate to be able to model achievement as a latent variable, thus taking advantage of latent modelling, such as directly correcting for measurement error.

[3] We included the covariates on the individual level (L1) only in order to control for selection effects and individual differences, rather than for additional compositional effects of these covariates. Adding all variables on the group level, however, did not change the structure of the results.

Table 1

Model Fit for all Models

No.	Model	*df*	χ^2	CFI	TLI	RMSEA
1.	Baseline model	41	244	>.99	.99	.02
	Correcting for unreliability and pre-existing differences					
2a	Measurement error	41	243	>.99	.99	.02
2b	Latent ACH factor	55	285	>.99	.99	.02
2c	Covariates	77	465	.99	.98	.02
	Combined model					
3	Model 2a+b+c	100	712	.99	.98	.02

Note. Measurement error = Baseline model additionally correcting for measurement error in the achievement variables; Latent ACH factor = Baseline model including mathematics achievement in Kindergarten as an indicator for the latent T1 achievement factor; Covariates = Baseline model including covariates;

Table 2

The Lag2 Direct and Total BFLPE and Peer Spillover Effect and their Effect Sizes for all Models

No	Model	BFLPE[a] T1-T2		BFLPE[a] T1-T3		
		Total	ES	Direct	ES	Total[b]
1	Baseline model	-0.34***	-0.15	-0.02	-0.01	-.11
2a	Measurement error	-0.36***	-0.16	-0.01	-0.01	-.96
2b	Latent ACH factor	-0.4***	-0.18	-0.08	-0.04	-.24*
2c	Covariates	-0.46***	-0.12	-0.1	-0.03	-.30**
3	Model 3 (2a+b+c)	-0.61***	-0.14	-.15	-0.04	-.34**
		PSE[a] T1-T2		PSE[a] T1-T3		
		Total	ES	Direct	ES	Total[b]
1	Baseline model	0.17***	0.13	-0.11***	-0.11	-.05
2a	Measurement error	0.11***	0.1	-0.03	-0.04	0.06
2b	Latent ACH factor	0.02	0.02	-0.23***	-0.22	-.22***
2c	Covariates	-0.08	-0.04	-0.11*	-0.06	-.18***
3	Model 3 (2a+b+c)	-0.2**	-0.1	-.06	-.04	-.22*

Note. [a] The compositional effects are calculated by subtracting the effect of individual achievement from the effect of school-average achievement. Total effects are the sum of all direct and indirect effects (for details on indirect effects see Supplemental Material). BFLPE =Big Fish Little Pond Effect; PSE = Peer spillover effect; T1 = Spring first grade; T2 = Spring third grade; T3 = Spring fifth grade; Direct = direct effect; ES = Effect size; Total = total effects; Measurement error = Baseline model correcting for measurement error ; Latent ACH factor = Baseline model including mathematics achievement in Kindergarten as an indicator for the latent T1 achievement factor; Covariates = Baseline model including covariates; * = $p < .05$;** = $p < .01$;*** = $p < .001$.

Table 3

The Effect of Covariates on all Variables in the Model 2c

Predictor	ACH1	ACH2	ACH3	MSC2	MSC3
Gender (Female = 1)	-0.06*** (-0.06***)	-0.06*** (-0.07***)	-0.02*** (-0.02***)	-0.17*** (-0.16***)	-0.05*** (-0.05***)
Age	0.16*** (0.17***)	-0.03*** (-0.03***)	-0.03*** (-0.05***)	-0.03** (-0.03***)	-0.03** (-0.03**)
Read	0.12*** (0.13***)	0.03*** (0.04***)	0.01 (0.04)	-0.01 (-0.01)	-0.02 (-0.02)
Book	0.05*** (0.05***)	0.01** (0.02**)	0.01* (0.01*)	-0.01 (-0.01)	0.03*** (0.03***)
Ethnic (Black =1)	-0.13*** (-0.14***)	-0.09*** (-0.09***)	-0.04*** (-0.05***)	0.06 (0.06)	0.09** (0.08**)
Ethnic (Hispanic =1)	0.07 (0.07)	0.05 (0.05)	0.01 (0.01)	0.04 (0.04)	-0.11 (-0.10)
Ethnic (Asian=1)	0.43 (0.45)	0.26 (0.28)	0.05 (0.05)	0.11 (0.10)	-0.36 (-0.33)
Ethnic (Other=1)	-0.46 (-0.48)	-0.28 (-0.29)	-0.04 (-0.03)	-0.14 (-0.14)	0.43 (0.4***)
SES	0.28*** (0.3***)	0.10*** (0.11***)	0.04*** (0.04***)	-0.02 (0.02)	0.01 (0.01)

Note. All Variables are modelled on L1. Unstandardized effects reported here (with standardized effects in brackets). ACH1 =students' mathematics achievement Spring first grade; ACH2 = students' mathematics achievement Spring third grade; ACH3 = students' mathematics achievement Spring fifth grade; MSC2 = mathematics self-concept Spring third grade; MSC3 = mathematics self-concept Spring fifth grade; Read = How often does student read outside of school; Book = Number of books at home ; Ethnic = Ethnical heritage; SES = Socioeconomic status; * = $p < .05$;** = $p < .01$;*** = $p < .001$

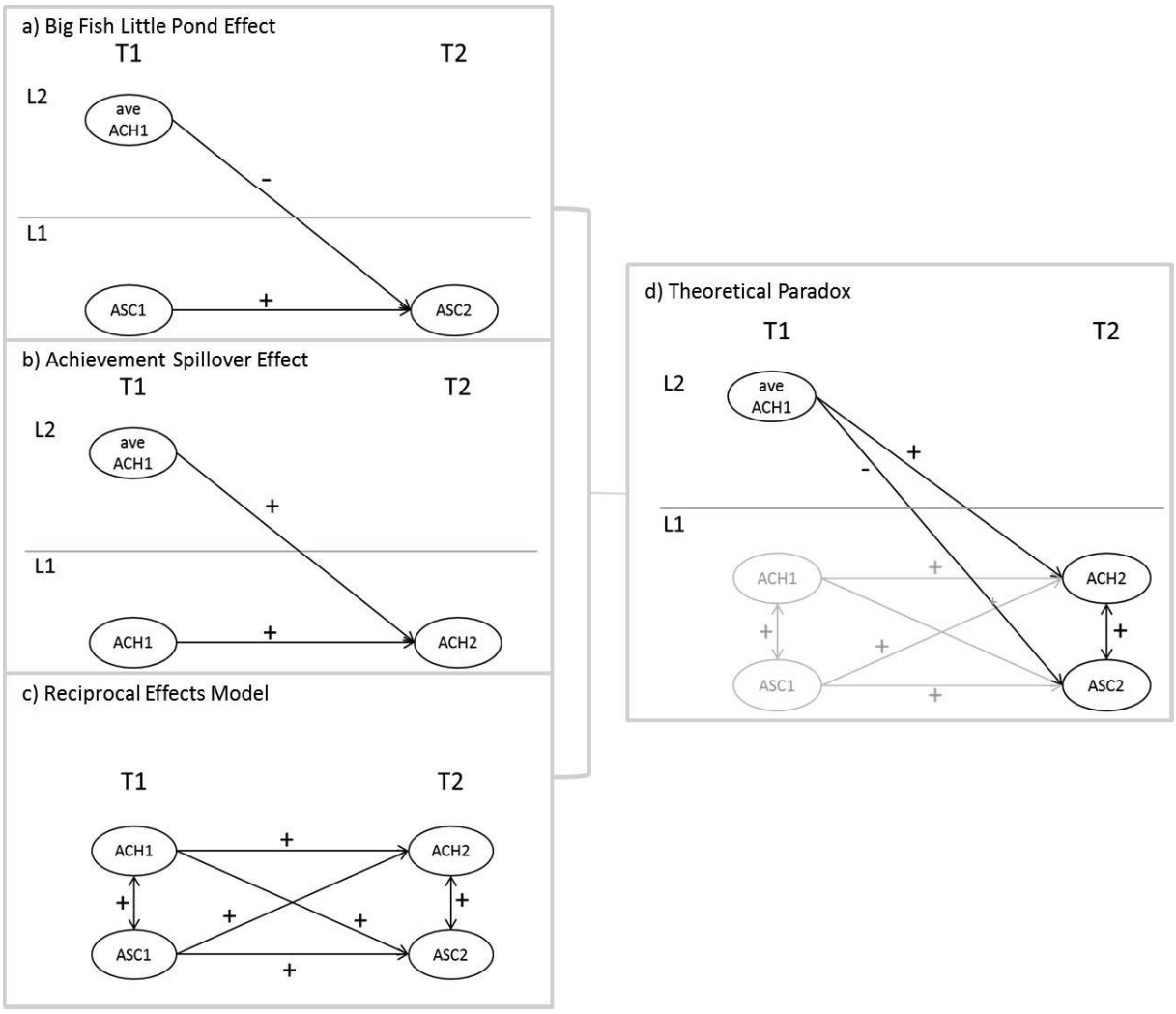

Figure 1.; a) Big-fish-little-pond effect; b) peer spillover effect; c) reciprocal effects model; d) the theoretical paradox after integrating a), b), and c). aveACH1 =aggregated students' academic achievement at T1; ACH1 =students' academic achievement at T2; ACH2 = students' academic achievement at T2; ASC2 = academic self-concept at T2; L1 = Individual student level; L2 = School level. T1 = first time wave; T2 = second time wave.

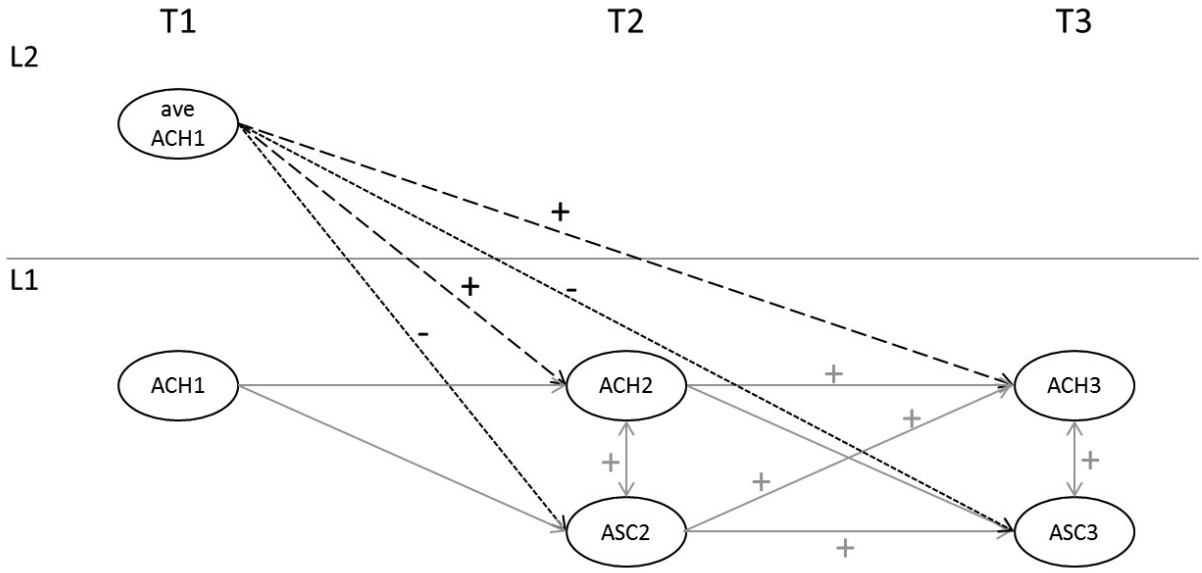

Figure 2. In this baseline model (Model 1) we included two longitudinal BFLPEs (Lag 1[T1-T2] and Lag 2 [T1-T3]) and simultaneously two peer spillover effects (Lag 1[T1-T2] and Lag 2 [T1-T3]). Thus, L2 achievement at T1 predicted L1 self-concept (BFLPE) and L1 achievement (Peer spillover effect). In addition this model included reciprocal effects at L1 in order to adequately model the longitudinal relationship of self-concept and achievement. Only theoretically relevant paths are shown here. aveACH1 =aggregated students' mathematics achievement Spring first grade; ACH1 =students' mathematics achievement Spring first grade; ACH2 = students' mathematics achievement Spring third grade; ACH3 = students' mathematics achievement Spring fifth grade; MSC2 = mathematics self-concept Spring third grade; MSC3 = mathematics self-concept Spring fifth grade; L1 = Individual student level; L2 = School level. T1 = Spring first grade; T2 = Spring third grade; T3 = Spring fifth grade; grey shaded lines = Reciprocal effects model; dotted line = Big-fish-little-pond effect; dashed lines = Peer spillover effect.