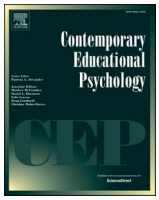
Contemporary Educational Psychology 84 (2026) 102441



Contents lists available at ScienceDirect

Contemporary Educational Psychology

journal homepage: www.elsevier.com/locate/cedpsych





Investigating the effects of class average achievement: attending a high-achieving class is neither beneficial for student achievement nor for academic self-concept

Theresa Dicke [a,*], Herbert W. Marsh [a], Philip D. Parker [a], Reinhard Pekrun [a,b], Jiesi Guo [a], Geetanjali Basarkod [a], Ioulia Televantou [c], Ziwen Teuber [d]

[a] Australian Catholic University, Institute for Positive Psychology and Education, North Sydney, Australia
[b] University of Essex, Department of Psychology, Essex, UK
[c] European University Cyprus, Cyprus
[d] University of Luxembourg, Luxembourg

ARTICLE INFO

Keywords:
Ability tracking
Big-fish-little-pond effect
Peer spillover effect
Class composition
Teaching quality

ABSTRACT

Research has shown positive effects of group-average achievement on individual achievement (peer spillover effect) but negative effects on academic self-concept (big-fish-little-pond effect; BFLPE), which is positively related to achievement. We examine these apparently paradoxical class composition effects in a longitudinal sample of more than 2000 German students followed over two school years in a tracked school system. Using multilevel models that control for school track, prior ability, socioeconomic status, gender, and two indicators of teaching quality (teacher enthusiasm and classroom management), as well as measurement and sampling error, we test the robustness of both effects.

Results show that the peer spillover effect is largely explained by pre-existing differences, such as prior ability. In contrast, the BFLPE became more negative after including covariates and remained so when controlling for teaching quality. The inclusion of teacher enthusiasm and classroom management only marginally changed either effect, indicating that, in this dataset, these two core instructional variables explained little of the variance in achievement or self-concept beyond pre-selection and peer composition effects. By distinguishing between compositional effects and instructional influences, our study contributes to a clearer understanding of how peer and classroom factors interact. Implications for future research include examining a broader range of teaching quality dimensions and exploring the conditions under which class composition may positively affect both achievement and self-beliefs.

1. Introduction

Certain educational assumptions have become so widely accepted that challenging them might appear counterintuitive. One such assumption is that sorting students by ability is beneficial for learning outcomes, a belief that has underpinned the adoption of tracking policies in many countries. As such, policies such as ability tracking, which reinforce schooling of students of similar abilities together, have been implemented in a large number of countries (Jerrim et al., 2015). Various outcomes have been found to be impacted positively by ability stratification (Hanushek et al., 2003). Specifically, there seems ample evidence for an (achievement) compositional effect, where group average achievement has a positive effect on individual students' achievement, over and above the positive effect of prior individual achievement (De Fraine et al., 2003; Stäbler, et al., 2017), referred to as "peer spillover effect" (Cooley Fruehwirth, 2013). Tracking policies are designed to create achievement-homogeneous classrooms, allowing for instruction to be tailored to students' ability levels (Hanushek & Wößmann, 2006). However, tracking systems do more than just sort students by prior achievement—they also systematically shape learning

* Corresponding author: Institute for Positive Psychology and Education, Australian Catholic University, 33 Berry Street, North Sydney, NSW 2060, Australia.
E-mail addresses: Theresa.dicke@acu.edu.au (T. Dicke), Herb.Marsh@acu.edu.au (H.W. Marsh), Phil.Parker@acu.edu.au (P.D. Parker), Reinhard.Pekrun@acu.edu.au, rp19684@essex.ac.uk (R. Pekrun), Jiesi.Guo@acu.edu.au (J. Guo), Geetanjali.Basarkod@acu.edu.au (G. Basarkod), I.Televantou@euc.ac.cy (I. Televantou), ziwen.teuber@uni.lu (Z. Teuber).

https://doi.org/10.1016/j.cedpsych.2025.102441

Available online 20 December 2025
0361-476X/© 2025 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

environments through differences in curriculum, peer composition, and teacher quality (Dustmann et al., 2017; Parker et al., 2021). This makes it difficult to isolate the effects of tracking from the potential benefits associated with being in classrooms with high-achieving students. Indeed, tracking is often justified based on the assumption that high-achieving peers will positively influence student outcomes, reinforcing the perception that tracking itself is beneficial (Zimmer & Toma, 2000). In this study, we test whether presumed benefits of tracking remain after controlling for pre-selection effects, which in our case include prior academic achievement (Year 4 grades), school track placement, and key demographic covariates such as gender and socio-economic status (SES). While tracking policies aim to reduce within-group variability by creating homogeneous classrooms, compositional effects refer to how the average achievement level of a class influences individual outcomes. One such effect, the *peer spillover effect*, describes how high-achieving classmates can raise individual achievement levels. Another effect, known as the *big-fish-little-pond effect* (BFLPE), occurs when being surrounded by higher-achieving peers lowers students' academic self-concept. Both effects are conceptually distinct from the intended instructional benefits of tracking (Dustmann et al., 2017). Another important mechanism not directly examined here is curriculum complexity. Higher tracks often deliver more advanced content and maintain higher academic expectations than lower tracks, which may influence outcomes independently of composition and instructional quality (Dustmann et al., 2017; Hattie, 2009; Kunter et al., 2013).

Proponents of tracking argue that homogeneity benefits both high- and low-achieving groups (Hattie, 2002; Dustmann et al., 2017; Hanushek & Wößmann, 2006). However, if peer spillover effects exist, tracking may inadvertently exacerbate disparities between students attending different tracks, as high-achieving students who typically attend higher tracks benefit from being surrounded by high-ability students, whereas low-achieving students attending lower tracks are surrounded by low-ability students. Imagine two students, A and B. A is placed in a high-achieving class with strong peers, while B is placed in a low-achieving class. If the peer spillover effect exists, A benefits from being surrounded by high-achieving peers, which further improves A's performance. In contrast, B, surrounded by lower-achieving peers, would not experience any improvement, but is more likely to be negatively affected by the low performance of B's environment. Over time, the gap in their academic outcomes grows wider, reflecting amplified disparities caused by tracking. In this study, we explore these compositional effects while accounting for instructional quality, thereby isolating the independent contributions of peer dynamics and class composition to student achievement and self-concept.

However, there is increasing doubt about the reliability and validity of the peer-spillover effect. Recent research using large samples, *meta*-analytic approaches, adequate statistical models, and/or strong controls for pre-existing individual differences suggests that a school system that reinforces homogenous class compositions does not have a positive effect on individual students' achievement (Dicke et al., 2018; Marsh, 1991; Televantou et al., 2015; Televantou et al., 2021) nor on average achievement or student aspirations in that system or country (Hanushek & Wößmann, 2006; OECD, 2013; Parker et al., 2016, Parker et al., 2018; Salchegger, 2016).

Moreover, research has provided strong empirical evidence of the negative effect of school-average achievement on academic self-concept (for an overview, see Marsh et al., 2008). This phenomenon has been termed the BFLPE (Marsh et al., 2008), where students with a given level of achievement have a lower academic self-concept in a class of high achievers (big pond) than in a class of low achievers (little pond). At the same time, academic self-concept is a strong predictor of later achievement (for an overview, see Seaton et al., 2015). Thus, when considering the proposed positive reciprocal relationship of academic self-concept and academic achievement (referring to the reciprocal effects model; Seaton et al., 2015), one can assume that being in a high-achieving class should be detrimental to individual students' achievement (Dicke et al., 2018).

If school and class composition (i.e., being schooled with high-ability students) are not the primary causes of the apparent differences, these advantages are likely because of a multitude of interrelated factors. Such factors may include differences in achievement and later occupational success between students who attended high-achieving versus low-achieving schools. While some might assume that class composition directly contributes to tracking benefits, prior research suggests that these advantages are largely driven by pre-existing individual differences rather than peer effects alone. Research has found prior achievement to be the strongest predictor of later achievement (e.g., Marks, 2010). It is well known that social class (or SES) shows a strong link to student outcomes, particularly achievement (e.g., Brantlinger, 2013; Parker et al., 2016, 2017).

Another common belief is that high-achieving or selective schools attract better teachers, which for most parents in itself warrants sending their children there. And indeed, beyond student characteristics and prior achievement, differences in teaching quality have been consistently identified as a key factor influencing student outcomes (Hattie, 2009). In tracked systems, schools and classrooms differ not only in peer composition but also in instructional quality, which may confound estimates of compositional effects. For example, high-achieving or selective schools often attract more experienced teachers, leading to instructional advantages independent of class composition (Rivkin et al., 2005). To disentangle compositional effects from broader school and instructional influences, we include two well-established indicators of teaching quality, that is teacher enthusiasm and classroom management, as control variables in our models (Frenzel et al., 2009; Korpershoek et al., 2016; Kunter et al., 2008).

In the study presented here we will expand existing approaches challenging the belief in the benefits of selection by ability, using (a) an unbiased aggregated measure of achievement assessed prior to selection into tracking, (b) controlling for other important demographic pre-selection effects such as SES and gender, and c) controlling for important variables of teaching quality.

1.1. Sorting by ability: research on compositional effects in education

We subsume the effects of aggregated peer, class, or school composition variables under the label compositional effects. Methodologically compositional effects are defined to be present whenever a given predictor variable at the aggregated group level (e.g., class level) has an effect on an individual outcome variable, controlling for the same variable at the individual level (Harker & Tymms, 2004). Here we investigate the effects of class composition of achievement – over and above the positive effect of prior individual achievement at the individual level – on:

(1) individual achievement. Much of the literature claims this effect is positive – the peer spillover effect (Cooley Fruehwirth, 2013; De Fraine et al., 2003; Stäbler et al, 2017)[1];
(2) academic self-concept, the perception of one's own academic abilities. Most research shows this effect to be negative – the BFLPE; Marsh et al., 2018).

Tracking will drive the compositional structure of a class through its sorting mechanism. Thus, research on class compositional effects is naturally related to research on the effects of tracking.

The consequences of tracking are subject to ongoing discussions (e.g., Vogl & Preckel, 2014). According to Hattie (2002), its effects can be grouped into two categories, achievement effects and equity effects. In the following, we will focus on explaining achievement effects.

[1] This means that achievement at an early wave acts as a predictor for later achievement (the same variable measured at a later point of time).

Regarding achievement effects, proponents of school tracking (or grouping students by ability) believe that allocating students to homogeneous learning environments leads to increased individual- and group-level achievement (Hattie, 2002; Dustmann et al., 2017). However, while tracking creates homogeneous classrooms, it also systematically alters educational environments through curricula, teacher quality, and peer achievement levels (Dustmann et al., 2017). Some studies suggest that students in classrooms with high achieving students may experience cognitive and academic benefits, including small positive effects on individual student IQ (Becker et al., 2012; Guill et al., 2017). More specifically, other studies have examined peer spillover effects, where class-average achievement may positively influence individual student performance (Becker et al., 2022; Retelsdorf et al., 2012).

While these findings may suggest benefits of being in a classroom with high-achieving students, they do not necessarily indicate that tracking itself is beneficial, as many of these effects diminish when controlling for pre-existing selection differences (Becker et al., 2022). In addition to considering pre-existing differences, it is critical to distinguish between compositional effects (such as the peer spillover effect) and tracking effects (homogeneous grouping) operate through different mechanisms.

A recent study by Becker et al. (2022) used an integrated data analysis approach to test the effects of class achievement on individual achievement in several German samples. Analyzing the peer spillover effect across five studies, the authors found, on average, significant positive effects of achievement class composition. Like other studies in this line of research, this compositional effect was reduced by adding strong controls, such as proxies for socio-economic status, to the model. Interestingly, the effect became even smaller after controlling for school track (of all 15 single class compositional effects on achievement, only five were still significant and positive, and one was significant and negative). This is likely because track serves as a strong proxy for prior achievement and other pre-existing differences that influence student outcomes. As a result, track indirectly captures cumulative academic performance, motivation, SES, and other cognitive and non-cognitive factors that influence later achievement (Borghans et al., 2016). Controlling for track, therefore, removes much of the variance in achievement that would otherwise be attributed to class-average achievement, suggesting that the initially observed peer spillover effect may be partly or entirely a phantom effect caused by unaccounted selection biases (Dicke et al., 2018; Televantou et al., 2021).

In contrast to these findings, other recent research using large samples, *meta*-analytic approaches, adequate statistical models, and/or

Table 1
Summary of studies referenced in our paper investigating peer spillover effects (achievement compositional effects) in various educational contexts.

Study	Sample	Predictor (group average)	Controls	Final peer spillover effect	Key findings
Becker et al., 2022	5 studies (range: 3,417–10,543 students)	Achievement	Measurement error, prior achievement, cognitive ability	Varies	Effects reduced with controls for measurement error and selection bias (out of 15 studies, nine became n. s., five remained positive and significant, one became negative and significant)
Dicke et al., 2018	14,985 students, 853 schools (US)	Achievement	Measurement error, gender, SES, parent variables	n.s.	Peer spillover effect reduced to below zero with controls
Cooley Fruehwirth, 2013	North Carolina public school students, grades 1996–2002	Achievement	Prior achievement	Positive	Adding prior achievement decreased the spillover effect
Guldemond & Bosker, 2009	1,125 students in 103 schools (1994 cohort), 941 students in 81 schools (1996 cohort)	Achievement	IQ, SES, prior achievement	n.s.	School compositional effects were not demonstrated; differences stemmed from proficiency levels already present at intake
Guill et al., 2017	8,628 German students from the academic and non-academic track and comprehensive schools	Intelligence	Prior achievement, migration background, cultural capital, parent schooling level	Positive	Students on academic tracks show greater intelligence score gains than students on other tracks
Hanushek et al., 2003	200,000 students in over 3,000 schools	Achievement	Various demographic and school-related variables	Positive	Sequential introduction of controls led to substantial changes in the magnitude and direction of peer effect estimates
Harker and Tymms (2004)	5393 students, 37 schools	SES	Prior achievement, ethnicity	n.s.	Positive effect of SES disappeared after including controls
Marks, 2015	40,143–40,273 students, 1289 schools	SES	Prior achievement	Small but negative	Positive school composition effect decreased after adding controls; small negative effect emerged
Marsh et al., 2023	Data from PISA 2000–2012 across multiple countries	Achievement, SES	Prior achievement, SES, measurement error	Small to negative	Peer spillover effects were small and often negative after controlling for measurement error, SES, and prior achievement. Results highlight SES impacts.
Nash, 2003	5,195 students in 362 schools (UK PISA data)	SES	Prior achievement, cultural capital, noise/disorder	n.s.	School composition effects were artifacts of statistical bias caused by omitted variables.
Stäbler et al., 2017	6,463 students, 151 schools	Achievement	SES, gender, language, track	Positive	Significant class-average effect controlling for prior achievement
Televantou et al., 2015 (Study 1)	19,059 students, 593 schools	Achievement	Measurement error	n.s.	Peer spillover turned insignificant with control for measurement error
Televantou et al., 2015 (Study 2)	1,694 students, 59 schools	Achievement	Measurement error	n.s.	Peer spillover turned insignificant with control for measurement error
Televantou et al., 2021	19,059 students, 593 schools	Achievement	Measurement error	Small and negative	Peer spillover turned even more negative with control for measurement error

Note. This table summarizes those studies referenced in our paper investigating peer spillover effects (achievement compositional effects) in various educational contexts. The *Sample* column describes the study populations, including the number of students and schools where applicable. The *Predictor (Group Average)* column indicates the group-level variable (e.g., achievement or SES) used to estimate the compositional effect. The *Controls* column lists covariates/pre-existing differences accounted for in the analyses, such as prior achievement, measurement error, or demographic variables. The *Final Peer spillover Effect* column indicates whether the compositional effect remained significant after including controls (n.s. = not significant). The *Key Findings* column provides a brief summary of the major conclusions, with a particular focus on how the inclusion of controls impacted the results.

strong controls for pre-existing individual differences suggests that a school system that reinforces children of similar ability being schooled together, for example through ability tracking, does not necessarily produce better learning outcomes at a national level. While tracking creates high-achieving and low-achieving classrooms by sorting students into groups of similar ability, it does not inherently improve achievement. More specifically, school-average achievement rarely has a direct positive effect on individual students' outcomes once selection effects are controlled (Dicke et al., 2018; Televantou et al., 2015; Televantou et al., 2021). Table 1 summarizes studies investigating the peer spillover effect that are referenced in the present article. It demonstrates that studies using more appropriate controls for pre-selection differences and measurement error tend to find smaller or non-significant peer spillover effects. This pattern is consistent, even though the peer spillover effect is not always eliminated. Overall, the findings show that robust controls are critical for isolating true compositional effects. Moreover, school systems that use some version of ability stratification, which leads to children of similar ability being schooled together and to less variance of achievement within schools, show lower overall levels of achievement, student aspirations, and student self-concept (Hanushek & Wößmann, 2006; OECD, 2013; Parker et al., 2016; Parkeret al., 2018; Parker, et al., 2021; Salchegger, 2016). Most commonly, observed positive differences in the achievement outcomes for higher-track students are due to the already existing differences, before any sorting of the students into groups of similar ability has come into effect (Dicke, et al., 2018; Parker et al., 2016; Parker et al., 2018).

Extensive support for the BFLPE generalizes across student groups, subject domains, academic self-concept instruments, and cultures (e.g., Marsh et al., 2023) and has been provided by experimental (e.g., Zell and Alicke, 2009) and longitudinal (e.g., Marsh et al., 2007; Keyserlingk et al., 2019) studies. Naturally, this negative effect implies that being in a low achieving class, school, or track leads to higher academic self-concept. This advantage however disappears after considering the side-effects of tracking such as signaling, where children are bound to a particular pathway (e.g., occupational in the vocational track) depending on their track (Buchmann & Park, 2009; Parker et al., 2018).

Interestingly, research has also found robust evidence of mutually reinforcing processes of individual academic achievement and academic self-concept in the reciprocal effects model (Valentine et al., 2004; Seaton et al., 2015). Integrating these findings results in a theoretical paradox where attending a school or class with high achievers is bad for a student's academic self-concept, but good for the student's individual achievement. Yet academic self-concept and achievement should be reciprocally and positively related (Dicke et al., 2018). This paradox may be resolved by findings that show the inconsistency of the peer spillover effect (e.g., Guldemond & Bosker, 2009; Hattie, 2002). Indeed, as mentioned above, significant positive peer spillover effects may result from poor methodology, such as not accounting for unreliability due to measurement and sampling error, and omitting important pre-selection effects. These methodological concerns suggest that the positive spillover effect on achievement is a *phantom effect* (Dicke et al., 2018; Harker & Tymms, 2004; Pokropek, 2015; Televantou et al., 2015). There are several straightforward techniques to account for measurement and sampling error, thereby reducing the bias in estimating compositional effects (e.g., by using doubly latent multi-level models; Marsh et al., 2009, Televantou et al., 2015). Several researchers have shown how measurement error can inflate compositional effects (Harker & Tymms, 2004; Pokropek, 2015; Televantou et al., 2015). For the negative BFLPE research has shown that the same methodological issues should downwardly bias the BFLPE (i.e., making it smaller), suggesting it is actually larger in reality (Parker et al., 2018).

The question of which other variables should be included in the models to avoid phantom effects is more complex. Such questions are dealt with more regularly in the realm of econometrics, where this issue is often termed as *the problem of endogeneity*. We will review the nature of endogeneity bias (or, more particularly, the omitted variable bias). To do so, we will use the research and language of econometrics.

1.2. The problem of endogeneity: omitted variable bias

The problem of endogeneity, although well-known in econometrics (Manski, 1993), is still largely ignored in applied research in other disciplines (Antonakis et al., 2014). Ewijk and Sleegers (2010) argue in their review and *meta*-analyses that while economic studies cautiously applied measures to obtain unbiased compositional effects, studies in other social sciences seem to have added covariates without deliberately aiming to reduce bias in the compositional effects. Put simply, endogeneity refers to the problems that arise when independent variables are correlated with the error terms (i.e., the unexplained variance) of dependent variables. Such a correlation can be caused by omitted variables not included in the analysis. In a regression equation of the form: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$,where β_0, β_1, $\beta_2, \cdots, \beta_k$ are parameters of interest and ε is a random error term, it is assumed that y is endogenous. This means it depends on the explanatory variables x_1, x_2, …x_k. In contrast, the x variables are assumed to be exogenous variables, meaning they are unrelated to the error term of the dependent variable (Wooldridge, 2015; Pearl, 2009). The problem of endogeneity and therefore potentially biased estimates due to omitting variables arises when at least one of the x variables in the model is related to the error term (Wooldridge, 2015). Econometrists have discussed several reasons for endogeneity of predictors, such as omitted variables, simultaneity (reversed causality), common method variance, etc. (Antonakis et al, 2014). The major cause however, seems to be that of omitted variable bias (Duncan et al., 2004), with all other causes in fact being a special case of omitted variables (Ebbes et al., 2005). Indeed, in their extensive review and *meta*-analysis Ewijk and Sleegers (2010) found that omitted variable bias seems to be the major issue in research on peer compositional effects (Duncan et al., 2004; Hanushek et al., 2003).

1.2.1. Bias through omission of important preselection variables

In research on compositional effects, the implications of omitting important variables are dire and can result in biased effects of the group level predictor variable due to pre-existing differences whereby variance in the outcome could in fact be explained by selection effects (Dicke et al., 2018; Hanushek et al., 2003; Harker & Tymms, 2004; Hauser, 1970; Marks, 2015; Nash, 2003). Group effects may be significant not because of group processes but because the individuals who make up the group were different to begin with. Thus, Hauser (1970) cautioned not to interpret residual differences between groups as social processes, as they might be explained through the correlation of the residuals with relevant individual student-level predictors. Among these predictors are most likely selection effects based on prior ability (achievement or IQ) or SES, that is, all students at a school in a certain neighborhood coming from a high or low SES background, which are usually associated with higher or lower achievement levels, respectively (Jerrim et al., 2015). After these predictors are accounted for, residual differences are reduced and possibly disappear altogether. Indeed, there is evidence that SES is related to ability, and that testing for the compositional effects of either SES or ability is confounded by the other variable (primary effect of social origin, see Boudon, 1974; Burns & Mason, 2002; Parker et al., 2016). For instance, research has shown that controlling for SES eliminated the positive peer spillover effect (Dicke et al., 2018), and others found the positive effect of group SES on individual achievement disappeared after including prior achievement (Harker & Tymms, 2004), or even turned slightly negative (Dicke et al., 2018; Marks, 2015).

1.2.2. Bias through omission of important resource variables

If it is not a school's composition that influences student achievement, it could be differences in the resources available to different schools. A common belief is that high-achieving or selective schools attract better teachers. This may be a sufficient incentive for parents to send their children to selective schools. Indeed, strong research found

the most influential school factor affecting student learning to be the teachers (Hattie, 2009). Hence, it makes sense to assume that a selection of elite teachers could be the crucial factor that could positively predict achievement (conditioned on pre-selection variables). Hattie's (2002) *meta*-analysis suggests, in explaining the lack of any effects of tracking, that it is the teachers who might adapt their attitudes and behaviors (see also Friedrich et al., 2015) based on peer composition, more than the composition that affect student achievement. Hence, it is critical for research to determine the unique compositional effect controlling for teaching quality.

We included teacher enthusiasm and classroom management as control variables to account for differences in instructional quality across classrooms. These variables were selected based on prior research showing robust links between classroom management and student achievement (Korpershoek et al., 2016) as well as between teacher enthusiasm and students' learning outcomes (Keller et al., 2016). Including these constructs allows us to control for selected aspects of classroom instruction, helping to ensure that any observed compositional effects are not merely reflections of differences in these instructional processes.

Teacher enthusiasm is one important indicator of teachers' attitude (for an overview see Keller et al., 2016). Teacher enthusiasm can be defined as nonverbal positive emotional experience of excitement by the teacher expressed as lively and engaging instructional behavior (Keller et al., 2016; Kunter et al., 2008). Frenzel et al., (2009) have shown that teacher enthusiasm is a mediator of the effect of teacher enjoyment on student enjoyment. Kunter et al. (2008) showed that teacher enthusiasm for teaching mathematics as perceived by students and self-reported by teachers was positively related to the teacher's instructional quality. Both factors should promote student achievement. Indeed, teacher enthusiasm has also directly been linked to achievement (Kunter et al., 2013; Keller, Neumann, & Fischer, 2013).

The ability to manage or prevent classroom disturbances or student misbehavior is termed "classroom management skills" (Evertson & Weinstein, 2006). Naturally, effective teaching and learning are not possible in poorly managed classrooms (Dicke et al., 2015). Meta-analyses show how classroom management strategies affected teachers' quality of instruction (Kunter et al., 2007) and, as a result, students' achievement (Hattie, 2009; Seidel & Shavelson, 2007; Wang, Haertel, & Walberg, 1993). Hence, we also include the ability to maintain an orderly classroom as a teacher variable in the present study.

1.3. The present study

In the present study, we used a large-scale dataset based on a sample of German year 5 (age: 11–12) and year 6 (age: 12–13) students. Rather than testing tracking effects per se, our study focuses on peer spillover effects and the BFLPE—two distinct compositional effects that emerge in different learning environments. We specifically investigate whether the assumed benefits of being in a classroom with high-achieving students due to the peer spillover effect persist once pre-selection effects, measurement error, and other confounds are accounted for. As we are using math ability as a predictor for compositional effects, we use math achievement and math self-concept as outcomes. Using doubly latent multilevel modeling (Dicke et al., 2018; Televantou et al., 2015; Televantou et al., 2021), we do this by ruling out effects of measurement and sampling error. We then investigate the effect of endogeneity caused by omitting important variables on estimates of the peer spillover effect. We focus on preselection effects by including SES (and gender) as covariates. Other potential extra-compositional effects in favor of classrooms or schools with high-achieving students and tracking could be resources such as teaching quality. Thus, in the present study we extend and complement prior research on the validity of compositional effects (e.g., Dicke et al., 2018) by including two important teacher variables in our model, that is, teacher enthusiasm and classroom management.

As an especially strong feature of this study, we base our class-aggregated achievement measure on achievement measured prior to selection into German school tracks by year four math grade (please see Supplemental Materials for a detailed description of how we modelled tracking and pre-selection Year 4 achievement and a rationale for the approach). Data from students enrolled in the German school system is particularly interesting for our investigation due to the early (Year 5) and explicit tracking system. In the Project for the Analysis of Learning and Achievement in Mathematics (PALMA) study, track allocation was determined by Year 4 grades, which represent a high-stakes measure of prior achievement in the German educational context. These grades are particularly salient because they serve as the primary selection mechanism for entrance into secondary school tracks, such as 'Realschule,' 'Hauptschule,' and 'Gymnasium.' Track, therefore, reflects a dual role: as a pre-treatment variable determined by prior achievement (Year 4 grades) and as part of the broader treatment, encompassing differences in teacher quality, curricula, and classroom climate. By conditioning our models on Year 4 grades, we are better able to control for selection effects and isolate the compositional effects of peer spillover and the BFLPE (please see Supplemental Material for more details on the German school system). Due to inherent disadvantages being maximal exacerbated by early ability tracking (Parker, et al., 2016; Dustmann et al., 2017), demonstrating any positive effects to be a myth would be even more important. In summary, our hypotheses are as follows:

Hypothesis 1. Adding important pre-selection variables (such as SES, gender and prior achievement) will reduce bias in the compositional effects resulting in a less positive peer spillover effect and an even more negative BFLPE. We note however that it is not possible to control for all pre-existing differences in non-experimental research, and that any unexplained residual variance is likely to bias results in favor of groups with initially brighter students.

It is additionally important to control for gender despite this not being a selection effect per se. Research (e.g., De Fraine et al., 2007) has shown stereotypical gender differences for math self-concept particularly. While mathematics achievement is very similar for female and male students, self-beliefs in math ability are lower in female than male students (e.g., Guo et al., 2015).

Hypothesis 2. Controlling for important resource variables as covariates (such as teacher enthusiasm and classroom management), in addition to the pre-selection variables (prior achievement variables, gender, and SES), will result in an even less positive peer spillover effect and an even more negative BFLPE compared to a model not controlling for these resource variables.

Thus, in the present study we extend and complement prior research on the validity of compositional effects (e.g., Dicke et al., 2018) by including two important indicators of teaching quality in our final model, that is, teacher enthusiasm and classroom management.

2. Method

2.1. Sample

Data for the present study come from a German large-scale longitudinal study that sought to investigate the development of math achievement along with its determinants (see Pekrun et al., 2007, 2019, 2023; Sakaki et al., 2025). The study was conducted in secondary schools in the German federal state of Bavaria. The German Data Processing Center (DPC) of the International Association for the Evaluation of Educational Achievement (IEA) was responsible for obtaining the samples and also for survey administration. Student samples ($N = 2070$; 50 % girls; mean age = 11.7 in Year 5; $SD = 0.7$) were drawn so that they were representative of the typical student population in the state of Bavaria in terms of student characteristics, geographic location, and SES. The sample comprised 83 classrooms from 42 schools. The sample

also included roughly equal numbers of students in the high achievement (Gymnasium: 37 %), middle achievement (Realschule: 30 %), and low-achievement (Hauptschule: 33 %) tracks. Self-report surveys and a mathematics achievement test were administered in a paper-and-pencil format to students in their classrooms by trained external test administrators at the end of each school year, from in Year 5 through to Year 10. The present study focuses on the first two of these time waves (Year 5 and Year 6) where class composition and teachers remained constant (there were no instances of changing class compositions or a change of teacher). This is important for our analyses of the effects of class composition. In addition, PALMA also includes participants' school grades from their last year of primary school (Year 4). Grades and demographic information were retrieved from school documents.

The study received Institutional Review Board approval from the Bavarian State Ministry for Education, Science, and the Arts (reference III/5-S4200/4-6/68 908). Participation in the study was voluntary, and parental consent was obtained for all students. The student participation rate was very high, at 91.8 %. Surveys were depersonalized to ensure participant confidentiality.

2.2. Measures

2.2.1. Math achievement

We assessed students' math achievement from three different sources. The first was school math grades from Year 4, which was a single-item end-of-year final grade. We re-coded the grades such that higher scores reflected better grades. As the grades were obtained directly from school documents, the data were error-free, except for the possible human error of copying. Results of a study by Marsh et al. (2018) indicate a high predictive and retest reliability of school grades (see Supplementary Materials for more detail). The second source was the school track whereby we identified which students attended Gymnasium, Realschule, and Hauptschule in Year 5. Coding Gymnasium = 1, Realschule = 2, and Hauptschule = 3, we modeled school track as a continuous variable. The third source was the PALMA Mathematical Achievement Test (vom Hofe et al., 2005), which was used to obtain standardized achievement test scores in math in Years 5 and 6. Using both multiple-choice and open-ended items, this test measures students' modeling and algorithmic competencies in arithmetic, algebra, and geometry. In each successive year, the test covered the same content areas, but the number and difficulty of the items increased in line with the year in school completed by non-repeating students. The obtained achievement scores were scaled using one-parameter logistic item response theory (Rasch scaling; Wu et al., 2007), and standardized in relation to Year 5 results (i.e., the first measurement point) to establish a common metric across the five waves. The reliability of item separation in item response theory scaling was 0.99 (Arens et al., 2017). Intraclass correlation coefficients (ICC[1]s) revealed a substantive amount of class-to-class variation in achievement; 0.54 in Year 4 (school grades); 0.39 in Year 5 (test scores) and 0.47 in Year 6 (test score).

2.2.2. Math self-concept

Math self-concept was measured in Year 5 and Year 6 using the PALMA self-concept scale (six items; see Supplemental Material for complete item list), e.g., "It is easy for me to understand things in math". Participants rated these items on a 5-point Likert scale ranging from 1 (not true at all) to 5 (absolutely true). Scale reliability and validity have been established in prior research, including factorial validity and evidence on relations with student achievement (Marsh et al., 2019; Pekrun et al., 2019). In the present analysis, the alpha estimates of reliability were high across both time waves (Year 5 $\alpha = 0.88$; Year 6 $\alpha = 0.89$). ICC[1]s revealed a small amount of between-classes variation in the self-concept factors; 0.06 at Year 5 and 0.07 at Year 6.

2.2.3. Teacher enthusiasm (perceived by students)

Students' perception of their teachers' enthusiasm was measured in Year 5 and Year 6, using the PALMA teacher enthusiasm scale, which was adapted from Marsh and Bailey's (1991) instructor enthusiasm scale (five items; see Supplemental Material for complete item list; e.g., "Our mathematics teacher teaches with enthusiasm"). Participants responded to these items using a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree), with higher scores indicating higher levels of perceived teacher enthusiasm.

Previous research has documented the psychometric quality of the scale (Frenzel et al., 2009, 2018). The alpha estimates of reliability were high at both timepoints (Year 5 $\alpha = 0.84$; Year 6 $\alpha = 0.85$). In order to examine the differences between classes for student ratings of teacher enthusiasm, we calculated the true ICC(1) for the latent factors (Muthén, 1991). Based on these factors, we calculated the ICC(2) to assess agreement within classes for the ratings of teacher enthusiasm (Bliese, 2000; Lüdtke et al., 2006; Snijders & Bosker, 1999). There were considerable differences in student ratings of teacher enthusiasm between classes (true ICC[1]: Year 5 = 0.206; Year 6 = 0.254). Agreement within classes, which can be interpreted as an index of reliability of the class-mean rating, was high (ICC[2]: Year 5 = 0.84; Year 6 = 0.87), as evidenced by ICC(2) values above 0.7.

2.2.4. Classroom management

Participants were presented with the following five items to measure students' perception of classroom management (see Supplemental Material for complete item list), e.g., "In mathematics class teaching is very often interrupted". The scale was adapted from Baumert et al.'s (1997) classroom management scale. Here we also used a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree), with higher scores representing better Classroom Management.

The alpha estimates of reliability were high in both Year 5 ($\alpha = 0.84$) and Year 6 ($\alpha = 0.84$). Results for between-class differences and within-class agreement were similar to those of teacher enthusiasm whereby there were considerable differences in teacher ratings between classes (true ICC[1]: Year 5 = 0.248; Year 6 = 0.240) and high reliability of the class-mean rating (ICC[2]: Year 5 = 0.87; Year 6 = 0.86).

2.2.5. Covariates

Students' gender (girls = 0 and boys = 1) and SES served as covariates. SES was assessed using the Erikson-Goldthorpe-Portocarero classification which was administered to parents (Erikson et al., 1979). The Erikson-Goldthorpe-Portocarero consists of six ordered categories of parental occupational status, with higher values representing higher levels of SES.

2.3. Statistical analysis

Using Mplus 8 (Muthén & Muthén, 1998–2010, Version 8), we conducted a series of doubly latent multilevel models. All models were analyzed as multi-level models to account for the nested structure of the data (students nested within classes), thus rendering standard errors that are corrected for nestedness: Level 1 (L1) = individual student, Level 2 (L2) = class. Moreover, in all our models we applied latent variables, thereby correcting for measurement error in our math self-concept, perceived teacher enthusiasm, and classroom management measures (see Dicke et al., 2018) (For a Mplus code of the final model please see Supplemental Material). In the present study, *pre-selection variables* refer to student characteristics that capture pre-existing differences before measuring our main outcomes. These include prior achievement (Year 4 mathematics grades), school track, SES, and gender. Of these, prior achievement and track are incorporated from the baseline model onward to account for the most direct forms of academic selection in the German tracked school system, whereas SES and gender are added in later models as additional demographic covariates. This definition clarifies that our use of "pre-selection variables" is broader than the covariates entered in a given model step.

Our compositional achievement variable was also modeled as a

latent variable based on latent aggregates of Year 4 grades (on L1 & L2) and track (on L2 only) in Year 5. This means Year 4 grades and track formed the indicators for the Year 5 latent ability variable. As track at Year 5 is a pure L2 variable with no variance within clusters on L1, we used Year 4 math grades and track as indicators on L2, but only Year 4 math grades as an indicator for L1 ability. We refer to these models which simultaneously control for measurement and sampling error as doubly latent based on Marsh et al.'s (2009) 2 × 2 taxonomy of multilevel models.

In Mplus, we used the robust maximum likelihood estimator (MLR), which is robust against any violations of normality assumptions. As has been done by other research (Dicke et al., 2018), we used the model constraint option in Mplus to calculate all compositional effects and their standard errors. We calculated compositional effects based on class composition (Marsh et al., 2014). The compositional effects are calculated by subtracting the effect of individual achievement from the effect of class-average achievement. These effects are analyzed independently of the instructional benefits claimed by proponents of tracking policies, focusing instead on achievement composition as a separate driver of achievement and self-concept differences. We estimated effect sizes employing the most conservative effect size measure ES3 as suggested by Marsh et al., 2009, which uses the total variance (varL1 + varL2) of achievement. For clarity of presentation of our results, we focus on direct paths of the longitudinal compositional effects.

2.3.1. Model sequence

In the present study, we present an a priori series of sequential models:

1. Baseline model (Model 1): On L1, included mathematics ability measured by Year 4 grades (prior achievement) as a predictor. On L2, included class-average Year 4 math grade and school track (academic placement in Year 5) as indicators of a latent ability variable. These two variables (prior achievement and track) are part of our definition of *pre-selection variables*, which capture pre-existing academic differences before the main outcomes. This latent ability variable predicted L1 self-concept (BFLPE) and L1 achievement (peer spillover effect) in Year 5 and 6
2. Model 2: Added further pre-selection variables in the form of demographic covariates: Gender on L1, and SES on both L1 and as a latent aggregate on L2 (see Marsh et al., 2023). This step allowed us to control for both academic and demographic pre-existing differences
3. Model 3: Based on model 2, additionally included teaching quality variables (teacher enthusiasm and classroom management) as time-varying covariates in Year 5 and 6, on L1 and L2. These were entered to control for instructional influences in the same year as the outcomes
4. For clarity, we added sensitivity analysis. Year 4 grades are the most critical prior measure of achievement to control for, as they represent the full variance of achievement in the student population in contrast to grades in later years, where the German tracking system affects variance. Our primary models condition on Year 4 grades and track to account for selection effects, allowing us to isolate compositional effects such as the peer spillover effect and the BFLPE

However, we estimated three additional models. One used class composition based only on Year 4 grades (no track), one used class composition-based Year 5 standardized test results as predictors, and the last included gender on L1 and L2. The first two models did not include track as a latent indicator of initial achievement. We estimated these models to address (1) the dual role of track as both a pre-treatment measure of prior achievement and a component of the treatment itself – by comparing models, and (2) the robustness of our findings relative to alternative specifications, (see Supplemental Materials Models S1, S3, and S4 respectively).

2.4. Model fit and missing data

All models were assessed using the following goodness-of-fit criteria, which were chosen as they are relatively independent of sample size (Marsh et al., 2004): the root mean square error of approximation (RMSEA); the Tucker-Lewis index (TLI); and the comparative fit index (CFI). For both the CFI and the TLI, population values vary along a 0-to-1 continuum, with a value of 0.90 being the minimum accepted threshold for goodness-of-fit. CFI and TLI values greater than 0.95 typically reflect excellent fits to the data. For the RMSEA, values smaller than 0.08 are considered acceptable model fits, while values <0.06 reflect good fits to the data.

At the time of the first assessment (Year 5), the sample size was 2070 students. The school context, combined with the longitudinal study design, inevitably results in some change in sample size due to some students, for instance, moving from or joining PALMA classrooms. This resulted in a sample size of 2059 in Year 6. Thus, a portion of the sample had missing data for at least one of the measurement waves. More specifically, the number of waves of data (Years 4–6) available for each student were: 3 (complete data; 52.7 %), 2 (16.7 %), and 1 (30.6 %). We did not exclude any cases because of missing data, but missing data was assumed to be missing-at-random (MAR) and handled using the full information maximum likelihood (FIML) approach which enabled us to use all available information for parameter estimation (Enders, 2010). For more details on our missing data approach, see Supplemental Materials.

3. Results

A correlation matrix for all variables included in the study can be found in Supplemental Material (Table SO2-3).

3.1. Baseline model

To test for uncorrected compositional effects, we modeled all four class compositional effects simultaneously in our baseline model (see Fig. 1). We included tracking and math achievement, measured by math grade in Year 4, as a latent variable, and mathematics test achievement and math self-concept at Time 2 (T2) and Time 3 (T3). This model showed a good fit to the data with CFI = 0.96, TLI = 0.96, RMSEA = 0.04 (see Supplemental Material Table S1 for more details on all model fit indices).

All models include four longitudinal direct compositional effects (from Year 4 to 5 and from Year 4 to 6): Two relating class-average Year 4 math achievement to later individual student achievement (peer spillover effect), and two relating class-average Year 4 math achievement to math self-concept (BFLPEs). For detailed results of all effects, see Table 2.

The Lag 1 peer spillover effect (Year 4–Year 5) was significantly positive. The Lag 2 peer spillover effect (Year 4–Year 6) was also significantly positive (see Table 2).

The Lag 1 direct BFLPE (Year 4–Year 5) was significantly negative. The Lag 2 direct BFLPE (Year 4–Year 6) was also negative and significant. For information on the direct L1 and L2 effects that comprise the compositional effects, please see Supplemental Materials Table S3.

3.2. Compositional effects corrected for selection effects (H1; model 2)

Although the overall setup of Model 2 was similar to the uncorrected baseline model, we additionally included further selection effects in our model to correct the compositional effects for pre-existing differences. This model fit the data well with CFI = 0.96, TLI = 0.96, RMSEA = 0.04. L1 variables were controlled for by gender and SES (see Table 3 for results and an overview of which variables were predicted by covariates). L2 variables were controlled for SES. For an overview of predictions by these covariates see Table 3 (estimates for the covariates were similar for

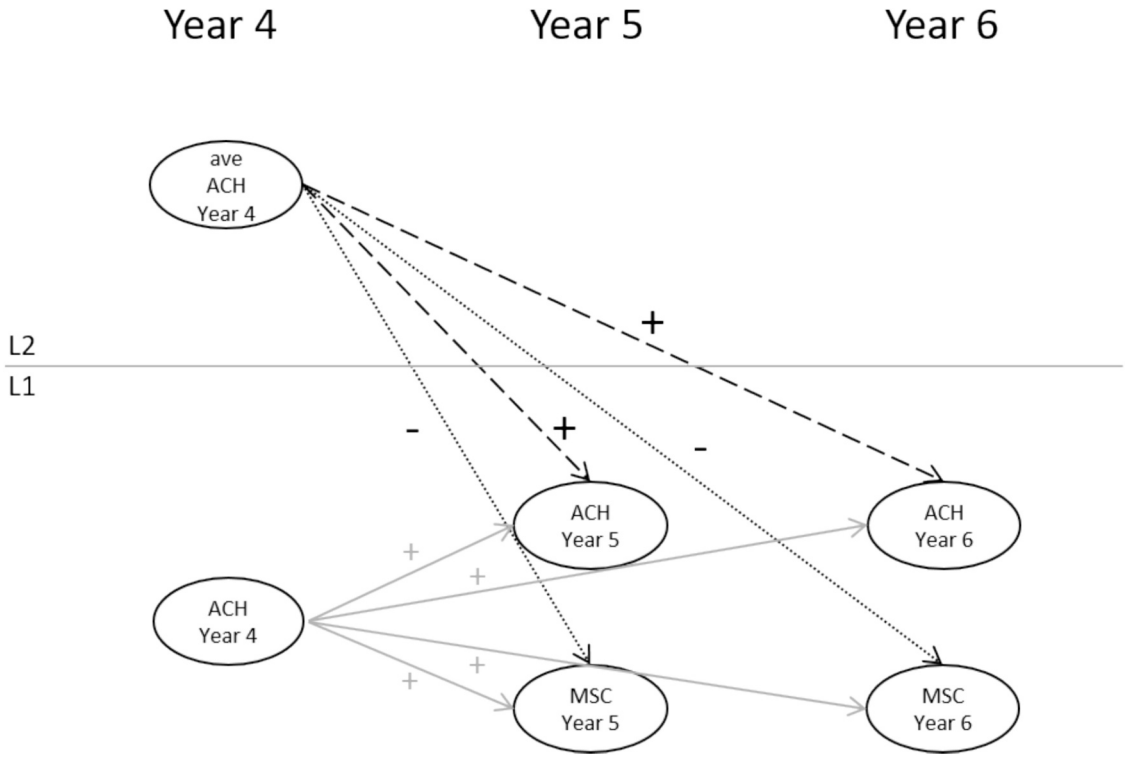


Fig. 1. In this baseline model (Model 1) we included two longitudinal BFLPEs (Year 4–Year 6 and Year 5–Year 6) and simultaneously two peer spillover effects (Year 4–Year 6 and Year 5–Year 6). Thus, L2 achievement in Year 4 predicted L1 self-concept (BFLPE) and L1 achievement (peer spillover effect). Only theoretically relevant paths are shown here, all variables on L1 are correlated if not otherwise depicted. aveACH year 4 = students' mathematics achievement measured by Year 4 mathematics grade (grades were reverse scored such that higher values indicated higher achievement) and track (higher values indicated higher track); ACH Year 5 = students' mathematics achievement measured by standardized test in Year 5; ACH Year 6 = students' mathematics achievement measured by standardized test in Year 6; MSC = mathematics self-concept; L1 = student level; L2 = class level; grey shaded lines = individual level effects; dotted line = Big-fish-little-pond effect; dashed lines = Peer spillover effect.

Table 2
Direct BFLPE and peer spillover effect and their effect sizes for all models.

		BFLPE[a] Year 4–5				BFLPE[a] Year 4–6			
		Coefficient	ES	Explained Variance %		Coefficient	ES	Explained Variance %	
No	Model			Within R^2	Between R^2			Within R^2	Between R^2
1	Baseline model	**−0.640(0.063)**	**−0.43(0.042)**	**0.14**	0.02	**−0.771(0.083)**	**−0.482(0.048)**	**0.14**	0.11
2	1 + Covariates	**−0.698(0.133)**	**−0.279 (0.058)**	**0.19**	0.06	**−0.840(0.123)**	**−0.310(0.063)**	**0.17**	0.12
3	2 + Teaching variables	**−0.722(0.112)**	**−0.293(0.055)**	**0.24**	**0.40**	**−0.893(0.109)**	**−0.339(0.060)**	**0.24**	**0.50**
		Peer Spillover Effect[a] Year 4–5				Peer Spillover Effect[a] Year 4–6			
		Coefficient	ES	Explained Variance %		Coefficient	ES	Explained Variance %	
				Within R^2	Between R^2			Within R^2	Between R^2
1	Baseline model	**0.226(0.056)**	**0.174(0.040)**	**0.22**	**0.80**	**0.367(0.058)**	**0.281(0.039)**	**0.21**	**0.86**
2	1 + Covariates	0.125(0.126)	0.065(0.066)	**0.23**	**0.81**	0.212(0.121)	0.115(0.069)	**0.22**	**0.88**
3	2 + Teaching variables	0.091(0.120)	0.047(0.062)	**0.24**	**0.83**	0.159(0.119)	0.087(0.068)	**0.22**	**0.90**

Note. [a] The compositional effects are calculated by subtracting the effect of individual achievement from the effect of class-average achievement (with standard errors in brackets). Significant effects with $p < 0.05$ in bold. School grades which were used for calculating achievement in Year 4 were reverse scored such that higher values indicated higher achievement. For track that was also used to calculate Year 4 achievement a higher value also indicated a higher track. BFLPE = Big-Fish-Little-Pond-Effect; ES = Effect size; Covariates = Socioeconomic Status and gender; Teaching Variables = student perceived classroom management skills and student perceived teacher enthusiasm; 1 + Covariates = Baseline model including covariates; 2 + Teaching Variables = Baseline model including covariates and teaching variables. Within R^2 = R^2 of outcomes on the individual level. Between R^2 = R^2 of outcomes on the class level.

Model 2 & 3 with differences of <0.01).

The Lag 1 peer spillover effect became less positive and insignificant after controlling for selection effects. The Lag 2 peer spillover effect likewise became less positive and non-significant. These results indicate that after taking into account selection effects the peer spillover effect disappears altogether. The Lag 1 BFLPE became *marginally* more negative and remained significant. The direct Lag 2 BFLPE also became *marginally* more negative and remained significant. This indicated that controlling for selection effects the BFLPE became even stronger (i.e., more negative; for details of these effects see Table 2). Effect sizes, however, did not increase in parallel with the coefficients. This is because adding covariates can alter the estimated latent variance components in doubly latent models, which changes the scaling of the effect size metric even when the underlying coefficient becomes larger.

Table 3
The effects of covariates on variables in the model 3.

	ACH Year 4	ACH Year 5	ACH Year 6	MSC Year 5	MSC Year 6
L1					
Gender	**0.120 (0.032)**	**0.111 (0.026)**	**0.060 (0.25)**	**0.237 (0.024)**	**0.201 (0.024)**
SES	0.031 (0.027)	0.060 (0.025)	0.077 (0.026)	**0.053 (0.021)**	**0.035 (0.026)**
Enthusiasm Year 5	–	**0.051 (0.022)**	–	**0.218 (0.025)**	–
Enthusiasm Year 6	–	–	**0.056 (0.020)**	–	**0.259 (0.030)**
Classroom Management Year 5	–	0.025 (0.021)	–	−0.021 (0.021)	–
Classroom Management Year 6	–	–	0.024 (0.023)	–	−0.047 (0.026)
L2					
SES	**0.823 (0.059)**	0.158 (0.144)	**0.274 (0.121)**	0.319 (0.308)	0.424 (0.222)
Enthusiasm Year 5	–	0.052 (0.054)	–	**0.519 (0.141)**	–
Enthusiasm Year 6	–	–	**0.111 (0.056)**		**0.569 (0.115)**
Classroom Management Year 5	–	**0.113 (0.055)**	–	0.111 (0.170)	–
Classroom Management Year 6	–	–	0.036 (0.057)	–	−0.159 (0.343)

Note. Standardized effects reported here (with standard errors in brackets). Significant effects with $p < 0.05$ in bold. ACH Year 4 = students' mathematics achievement measured by track and Year 4 mathematics grade (grades were reverse scored such that higher values indicated higher achievement; track was coded such that higher numbers indicated higher track); ACH Year 5 = students' mathematics achievement measured by standardized test in Year 5; ACH Year 6 = students' mathematics achievement measured by standardized test in Year 6; MSC = mathematics self-concept; L1 = student level; L2 = class level; SES = Socioeconomic status, Gender was coded as 0 = girls and 1 = boys.

3.3. Compositional effects corrected for selection effects and teaching quality (H2; model 3)

In addition to the covariates that control for selection effects (Model 2) we added controls for two teaching quality variables as time-varying covariates. Thus, we tested if perceived teacher enthusiasm and classroom management assessed in Year 5 and 6 predicted all variables within the respective time wave (see Table 3 for results and an overview of which variables were predicted by covariates). This model also fit the data well with CFI = 0.95, TLI = 0.95, RMSEA = 0.03. At L1, the individual deviation from the group mean of teacher enthusiasm significantly and positively predicted achievement and math self-concept, while the individual deviation from the group mean for classroom management showed only small and non-significant effects. At L2, both class-average teacher enthusiasm and class-average classroom management were only weakly and non-significantly related to achievement and math self-concept. For an overview of predictions by all covariates see Table 3. Both teaching variables showed moderate stability on L1 and high stability on L2 (see Supplemental Materials Table S4. All peer spillover effects became even less positive, but this reduction, however, was only small (see Table 2). The Lag 1 and Lag 2 BFLPEs only changed slightly, becoming more negative.

At the between level, the explained variance for the BFLPE in self-concept was nonsignificant in the baseline and covariate-only models, but significant in the full model including teaching quality; all other within- and between-level R^2 values were significant.

3.4. Sensitivity analysis

Given the high correlation between Year 4 grades, Year 5 test scores, and track, the sensitivity models showed patterns that were consistent with our primary analyses. Specifically, the BFLPE remained significant and stable across all models, while the initially positive peer spillover effect diminished and became non-significant with the inclusion of additional controls. These results further support the robustness of our findings. For more details, see Supplemental Materials Table S2, Models S1 and S3, respectively.

4. Discussion

The present study showed that the peer spillover effect is significant and positive only when important omitted variables are excluded in the model. Thus, expanding prior research (Dicke et al., 2018; Harker & Tymms, 2004; Hauser, 1970), our study emphasizes the importance of addressing the problem of endogeneity and hence, omitted variable bias in social sciences (Duncan et al., 2004; Ewijk and Sleegers, 2010).

Our findings suggest that while tracking policies aim to benefit all students through tailored instruction in achievement-homogeneous settings (Hanushek & Wößmann, 2006), the actual benefits of such policies are difficult to identify when compositional effects are disentangled from pre-existing selection factors. The reduction of the peer spillover effect's size to near zero suggests that the academic benefits often associated with classrooms with high-achieving students may largely reflect pre-existing differences rather than the influence of peer achievement per se. In contrast, the stability and even increase of the negative BFLPE highlights an adverse consequence of tracking: students placed in high-achieving classes often experience negative effects on their academic self-concept due to unfavorable social comparisons with their high-achieving peers. Together, these findings call into question whether tracking achieves its intended instructional goals or instead exacerbates inequalities in both academic outcomes and student self-beliefs.

4.1. Preselection variables account for most peer composition effects (Hypothesis 1)

Indeed, our research provided strong evidence that pre-existing differences, rather than the classes' peer composition of achievement, explain variance in individual achievement (Dicke et al., 2018; Marks, 2015; Nash, 2003). In addition, we theoretically assumed that a model with compositional effects based on "pre-treatment" variables (track and pre-selection math grades) would show even stronger effects than a model based on "post treatment" achievement. However, results of a supplemental model (see Supplemental Materials, Table S2 Model S1) that based the compositional effect on the Year 5 standardized test scores were very similar. That said, Year 5 is still only one year after the "treatment" tracking has occurred, research shows that the BFLPE increases over time if students remain in the same educational setting (for an overview see Marsh et al., 2007). Hence, we would still recommend using any "pre-treatment" variables to model compositional achievement effects. In situations such "pre-treatment" data is not available, data collected soon after treatment may be used.

SES proved to be one of the most important pre-selection variables in our model. In our model, after including SES and gender as controls, and while using prior achievement and track as a predictor, the peer spillover effect was largely reduced, similar to other studies with German data (Becker et al., 2022). More specifically, the peer spillover effect was almost zero and not significant. The BFLPE, however, turned slightly more negative after controlling for SES and gender. These findings highlight a concerning pattern: students in classrooms with high-achieving students report lower academic self-concepts, and the anticipated benefits to achievement—often associated with being in such environments—were not observed once pre-existing differences were

accounted for. Instead, any findings showing positive effects of membership in a classroom with high-achieving students on individual achievement are largely due to the omission of important pre-selection variables. It is important to note that our findings are based on one German sample in one specific German state (Bavaria) as opposed to e.g., Becker et al., (2022) who based their findings on 5 different samples and found an average remaining positive peer spillover effect. The advantage, however, of using a sample from this particular state lies in the fact that Bavaria has very strict cut-off criteria for the allocation to school-track (for more details see above and Bavarian Ministry of Culture and Education, 2020). In this sense, track is even more reflective of prior achievement here than in other German states. Nevertheless, further investigation into the role of track and the source of differences in spill-over effects is needed (Becker et al., 2022).

4.2. Teacher enthusiasm and classroom management have limited impact (*Hypothesis 2*)

Based on our review of important resource variables, we expected teacher enthusiasm and classroom management, two core aspects of teaching quality examined in this study, to have an impact on student achievement (Hattie, 2009; Kunter et al., 2013; Keller, et al., 2013; Muijs & Reynolds, 2002; Wang et al., 1993). The small Level-2 effects of these constructs on student achievement were surprising as our model setup should provide a "pure" effect of teaching (or perception of teaching) by controlling for pre-selection variables as well as the influence of class composition (Rivkin et al., 2005). This small teaching effect contrasts a number of other studies which tend to report larger effects (Hattie, 2009). One possible explanation for this is that the effect sizes in other studies may appear larger because they do not include these controls, specifically pre-selection variables and the influence of class composition. Furthermore, it may be even more challenging to distinguish between compositional effects (peers) and teaching variables (e.g., engagement, classroom management) than initially assumed (Decristan et al., 2017; Gnas et al., 2022). For example, there might be variables, such as classroom climate, which are a function of the group of students in the classroom, but not a compositional effect per se, more than they are an effect of the teacher who teaches the class (Marsh et al., 2008) and are highly related to variables such as classroom management and teacher enthusiasm. These results suggest that the specific teaching-related variables included in this study, namely *teacher enthusiasm* and *classroom management*, do not account for the observed compositional effects. This supports the conclusion that, at least in this sample, such instructional features do not explain the benefits often attributed to being in a classrooms with high-achieving students beyond what is explained by students' prior characteristics.

Results for the compositional effects after controlling for teacher enthusiasm and classroom management were in the hypothesized direction, with a peer spillover effect becoming even closer to zero and the BFLPE becoming more negative. The differences were, however, small, indicating that teacher effects did not make a big difference. This finding is in line with other research showing that the relation between classroom management and classroom composition effects might be small (Decristan et al., 2017). This is also related to the assumption that the class composition affects student outcomes indirectly through teacher perceptions and behaviors (Hattie, 2002; Friedrich et al., 2015).

4.3. Strengths, limitations, and implications for future compositional effect research

The present study is innovative in applying state-of-the-art approaches of economics and psychology to methodologically strong longitudinal models (doubly latent multilevel modeling) that are theory-driven. Furthermore, we included important achievement-related and resource-related covariates and based our aggregated achievement variable on pre-selection achievement.

It is important to note that our modeling approach using grades and track as indicators is not perfectly symmetric, as grades are pre-selection and pre-treatment, while track is part of the treatment. However, the decision about track is pre-treatment. Further, we use math grades, which is based on one subject and track is based on an overall assessment of the students' abilities. Nevertheless, we think the resulting variable is a strong measure of pre-existing differences given the tight coupling of math grade and the tracking decision. Supplemental analyses using only math grades as an indicator of the compositional effect revealed very similar results (Supplemental Materials Table S2, Model S3).

However, there are inevitably other omitted covariates. Considering them could control for additional pre-existing differences (e.g., migration status, geolocation) and available resources, which we did not include (Hanushek et al., 2003; Harker & Tymms, 2004; Hauser, 1970). Our study includes teacher enthusiasm and classroom management as key indicators of teaching quality, as these factors have been widely linked to student learning outcomes (Hattie, 2009; Kunter et al., 2013). However, we acknowledge that additional instructional factors, such as cognitive activation, instructional clarity, and teacher support (Praetorius et al., 2018), could further contribute to understanding how teaching quality affects compositional effects. Future studies should aim to incorporate a broader range of teaching-related variables to provide a more comprehensive analysis of teacher influences on student achievement and self-concept. In addition to compositional effects, high-achieving educational settings often offer many other benefits, such as advanced curricula (Dustmann et al., 2017), better instructional quality (Hattie, 2009; Kunter et al., 2013), and favorable classroom climates (Decristan et al., 2017). Even though our study isolates compositional effects from potential benefits such as advanced curricula, higher instructional quality, or favorable classroom climates, we recognize that these broader mechanisms have been shown in previous research to contribute to the advantages of high-achieving schools or tracks. These mechanisms were not examined in our study, and our findings should therefore not be interpreted as an evaluation of them. Rather, we emphasize the importance of distinguishing compositional effects from these other factors, which could benefit students regardless of their achievement level. While controlling for variables such as curriculum, resources, and school leaving certificates may be considered statistical overcontrol in practice, this approach is critical for investigating the unique contribution of compositional effects. By decomposing global variables like tracking, we aim to isolate the impact of peer spillover effects and the BFLPE, independent of other mechanisms that influence student outcomes. Further, by controlling for these variables, we remove many mechanisms that explain why attending high-achieving schools or tracks might be advantageous in practice, such as differences in curricula, instruction, and school leaving certificates. Consistent with our assumptions, including these in future studies, would be expected to make the peer spillover effect even smaller or negative, and the BFLPE even more negative.

In the present study, we investigated the effects of teacher enthusiasm in predicting student achievement in line with existing research (e.g., Frenzel et al., 2009; Kunter et al., 2008). It is reasonable, however, to assume that this relationship is reciprocal in nature (similar to Frenzel et al., 2018), as better student achievement could affect (the perception of) teacher enthusiasm and also classroom management. Future longitudinal studies that focus on teaching quality in predicting achievement should consider reciprocal relationships of these variables. Although not measured here, curriculum complexity is an important contextual factor alongside peer composition and instructional quality. Differences in the breadth, depth, and pace of curriculum across tracks are well-documented (Dustmann et al., 2017; Hanushek & Wößmann, 2006) and likely contribute to the advantages often associated with higher tracks. Its absence does not affect our estimation of compositional effects, but future research should investigate the effects of curriculum complexity explicitly.

In the present study, we utilised multi-level doubly latent longitudinal structural equation modelling to account for endogeneity, sampling error and measurement error in estimating compositional effects. There are, however, other methodological approaches that could strengthen causal claims in this research area. One promising approach is the student fixed-effects strategy suggested by Loyalka et al. (2018) that controls for all student, classroom and school characteristics as well as a large number of observable subject-specific characteristics. Using a large-scale cross-national database, Loyalka et al. (2018) provide a more rigorous test of the BFLPE across diverse educational settings than conventional approaches to compositional effects (OLS regression; Loyalka et al. (2018)). This approach could also be used to investigate the peer spillover effect to enable making similar strong causal inferences."

Additionally, natural experiments, such as policy changes leading to exogenous variation in school composition, offer an alternative avenue for causal inference (Dustmann et al., 2017). Difference in differences approaches to compare policy changes that only affect local areas could also be utilized. Finally, instrumental variable (IV) approaches, where exogenous factors (e.g., percentage of social housing in a school area) serve as instrumental variables in analyzing compositional effects. This could help isolate the causal impact of peer composition from correlated school characteristics. Combining these methods with the doubly-latent multilevel modeling used in this study could provide an even more robust understanding of the peer spillover effect and its interplay with tracking policies.

The current research investigated effects on math achievement and math self-concept. Consequently, the interpretation of results and any conclusions can only be made for the math domain. However, the BFLPE is one of the most robust findings in educational psychology and has been found for various other domains as well (Guo et al., 2018; Marsh et al., 2008). The peer spillover effect and the decline of estimates after utilizing stronger methodology and control variables have also been shown over multiple domains (e.g., Becker et al., 2022). Further, phantom effects and omitted variable bias are methodological issues that will occur irrespective of the chosen achievement domain. As such, the current findings are likely to be generalizable to other academic subjects. This, however, is an interesting research question for future research.

This study was conducted with a sample of German students which makes results particularly interesting due to the early implemented and explicit tracking system in which inherent disadvantages of ability tracking are maximally exacerbated (Parker, et al., 2016; Dustmann et al., 2017), making this line of research even more important. Among these disadvantages are delayed effects like the effects of signaling on students' occupational possibilities, but also immediate effects on students' well-being. For example, for a group of high-achieving kids, which should theoretically benefit the most from tracking, Becker et al. (2014) showed detrimental effects of early transition into tracked schools on self-concept and school anxiety as well as a lack of increase in the development of peer relations compared to a matched group of students who did not enter tracking early.

5. Policy and practice implications: supporting achievement and self-concept

Our findings also have implications for educational policy and systems. First, the robustness of the BFLPE (Marsh et al., 2007) highlights the importance of addressing students' motivational and emotional well-being in high-achieving contexts. Policy-makers and schools should implement interventions that promote positive self-concept for students, particularly in high-achieving tracks and contexts. Hence, we suggest that policies should aim to balance academic success with motivational and emotional well-being by promoting positive self-views, particularly in competitive environments (Basarkod et al., 2023). Tailored or individualised feedback strategies and self-concept interventions (Hörsch et al., 2025) can help buffer the negative effects of social comparison.

Second, the reduction in peer spillover effects after controlling for pre-selection variables calls into question the assumption that tracking inherently benefits all students. This is in line with research (Matthewes, 2021; Parker et al., 2016) that suggests the apparent benefits of (early) tracking are largely driven by systemic disparities, such as unequal access to high-quality instruction and resources. Enabling equal access to resources like advanced curricula and better instructional quality across tracks could ensure that all students benefit, regardless of their school's track.

From a policy perspective, our results suggest that relying on compositional effects alone, without addressing broader systemic disparities, may reinforce educational inequalities (Hanushek & Wößmann, 2006; Parker et al., 2018). Policymakers should critically assess tracking policies and consider complementary measures, such as targeted resource allocation and support systems, to mitigate the disadvantages faced by lower-track students.

6. Conclusion: small peer spillover, robust negative BFLPE

Taken together, our results enrich the evidence on compositional effects and emphasize the important question: Are there benefits of being schooled amongst high achievers? According to the present results, we can draw the following conclusion: For countries with tracked and socio-economically segregated education systems (here: Germany), using adequate statistical modelling and reducing endogeneity bias results in a non-significant and very small peer spillover effect as well as a negative BFLPE. Our analysis distinguishes between compositional effects, such as the peer spillover effect and BFLPE, and other benefits often offered in high-achievement settings, such as instructional quality, better curricula, and better school resources. This separation is critical for understanding the specific role of peer composition while acknowledging the other broader mechanisms that benefit students in higher-achieving tracks. The teaching variables included in this study, that is, teacher enthusiasm and classroom management, did not play a major role in either compositional effect. We conclude that regarding individual achievement as an outcome, class-average achievement does not appear to have a positive impact over and above the effects of pre-existing selection variables and teaching quality. Put simply, our findings suggest that the academic benefits often attributed to classrooms with high-achieving students are largely explained by students' prior ability, SES, and gender, rather than the classroom context itself.

The BFLPE will lead to a decrease in self-concept for (mostly advantaged) students in high-achieving classes and an artificially high self-concept for (disadvantaged) students in low achieving classes (Parker et al., 2017). This effect might appear to balance out pre-existing inequity as the boosted self-concept of disadvantaged students should in turn be positively related to aspirations (Guo et al., 2015) and achievement (Dicke et al., 2018). However, research by Parker et al. (2017) shows strong evidence that the process is far more complicated and has other unintended consequences. Parker et al. (2017) argue that school tracking will concurrently enable signaling processes (Buchmann & Park, 2009), where attending and thereby being associated with a certain school (type) will lead to a preconceived idea of a student's ability irrespective of what that might be. In case of a disadvantaged student, this may lead to being labeled as a low achiever and being withheld opportunities that would be available to more advantaged peers, based merely on the type of school the disadvantaged student attends (Parker et al., 2017; Buchmann & Park, 2009).

Moreover, research has demonstrated that several measures of educational inequity (including explicit and implicit school tracking) are negatively related to countries' average achievement (Hanushek & Wößmann, 2006).

Using data from five cycles of Programme for International Student Assessment (PISA) assessments (PISA 2000 to PISA 2012) across 27 OECD countries, Parker et al. (2018) demonstrated that countries with

greater ability stratification (more tracking) had lower average student achievement. Additionally, countries that increased ability stratification over time showed declining achievement levels, particularly for low- and average-achieving students. These findings highlight how country-level inequality that is driven by stratification can be harmful for student excellence based on achievement.

Matthewes (2021) further underscored the adverse impacts of early tracking systems, especially on equity and efficiency in the German context. His findings revealed that prolonged comprehensive schooling significantly benefits low-achieving students without harming high achievers. This supports the argument that the perceived advantages of high-achieving settings often stem from systemic disparities rather than compositional effects alone.

From a policy perspective, these findings suggest that the benefits of high-achieving settings often derive from systemic advantages, such as better resources and curricula, rather than compositional effects alone (Dustmann et al., 2017; Hattie, 2009). Policymakers should focus on expanding access to these benefits for students across all tracks. For example, ensuring that lower-achieving tracks have access to the same instructional quality and advanced curricula as higher tracks could reduce disparities while maintaining the positive aspects of tracking systems (Becker et al., 2022; Decristan et al., 2017). While our study does not evaluate tracking policies per se, our findings indicate that the academic benefits often attributed to classrooms with high-achieving students are largely explained by pre-existing differences rather than peer composition.

We conclude that the benefits of attending high-achieving classes or schools are questionable when compositional effects are isolated, and other factors like instructional quality, curricula, and school leaving certificates come into play. Attending such schools may also harm students' self-beliefs.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used ChatGPT, Wordtune, and Grammerly in order to improve writing and reduce typos. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

9. Code availability

In Supplemental Material.

CRediT authorship contribution statement

Theresa Dicke: Writing – review & editing, Writing – original draft, Visualization, Software, Project administration, Methodology, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Herbert W. Marsh:** Writing – original draft, Methodology, Funding acquisition, Conceptualization. **Philip D. Parker:** Writing – original draft, Visualization, Methodology, Formal analysis, Conceptualization. **Reinhard Pekrun:** Writing – original draft, Methodology, Funding acquisition, Data curation, Conceptualization. **Jiesi Guo:** Software, Methodology, Formal analysis. **Geetanjali Basarkod:** Writing – review & editing, Writing – original draft, Methodology. **Ioulia Televantou:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Ziwen Teuber:** Writing – review & editing.

Funding

This work was supported by the Universities Australia (UA) −Deutsche Akademische Austauschdienst (DAAD) Australian-German Joint Research Co-operation Scheme.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cedpsych.2025.102441

Data availability

Data will be made available on request.

References

Arens, A. K., Marsh, H. W., Pekrun, R., Lichtenfeld, S., Murayama, K., & Vom Hofe, R. (2017). Math self-concept, grades, and achievement test scores: Long-term reciprocal effects across five waves and three achievement tracks. *Journal of Educational Psychology, 109*(5), 621–634.

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2014). Causality and endogeneity: Problems and solutions. *The Oxford Handbook of Leadership and Organizations, 1*, 93–117.

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 cross-national exploration with PISA 2018. *Scientific Studies of Reading, 27*(4), 375–392.

Bavarian Ministry of Culture and Education (2020). Übertritt und Schulwechsel in Bayern: So geht die Schulkarriere weiter. https://www.km.bayern.de/eltern/schularten/uebertritt-schulartwechsel.html.

Baumert, J., Gruehn, S., Heyn, S., Köller, O. & Schnabel, K.-U. (1997), *Bildungsverläufe und psychosoziale Entwicklung im Jugendalter (BIJU), Dokumentation*, [Educational careers and psychosocial development in adolescence (BIJU): Documentation]. Berlin: Max-Planck-Institute for Human Development.

Becker, M., Kocaj, A., Jansen, M., Dumont, H., & Lüdtke, O. (2022). Class-average achievement and individual achievement development: Testing achievement composition and peer spillover effects using five German longitudinal studies. *Journal of Educational Psychology, 114*, 177–197.

Becker, M., Lüdtke, O., Trautwein, U., Köller, O., & Baumert, J. (2012). The differential effects of school tracking on psychometric intelligence: Do academic-track schools make students smarter? *Journal of Educational Psychology, 104*, 682–699.

Becker, M., Neumann, M., Tetzner, J., Böse, S., Knoppick, H., Maaz, K., … Lehmann, R. (2014). Is early ability grouping good for high-achieving students' psychosocial development? Effects of the transition into academically selective schools. *Journal of Educational Psychology, 106*(2), 555.

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis.

Borghans, L., Golsteyn, B. H. H., Heckman, J. J., & Humphries, J. E. (2016). What grades and achievement tests measure. *PNAS Proceedings of the National Academy of Sciences of the United States of America, 113*(47), 13354–13359. https://doi.org/10.1073/pnas.1601135113

Boudon, R. (1974). *Education, opportunity, and social inequality. changing prospect in western society*. Neuwied: Luchterhand.

Buchmann, C., & Park, H. (2009). Stratification and the formation of expectations in highly differentiated educational systems. *Research in Social Stratification and Mobility, 27*, 245–267.

Burns, R. B., & Mason, D. A. (2002). Class composition and student achievement in elementary schools. *American Educational Research Journal, 39*, 207–233.

Brantlinger, E. (2013). *Dividing classes: How the middle class negotiates and rationalizes school advantage*. Routledge.

Cooley Fruehwirth, J. (2013). Identifying peer achievement spillovers: Implications for desegregation and the achievement gap. *Quantitative Economics, 4*, 85–124.

Decristan, J., Fauth, B., Kunter, M., Buettner, G., & Klieme, E. (2017). The interplay between class heterogeneity and teaching quality in primary school. *International Journal of Educational Research, 86*, 109–121.

De Fraine, B., Van Damme, J., Van Landeghem, G., Opdenakker, M. K., & Onghena, P. (2003). The effects of schools and classes on language achievement. *British Educational Research Journal, 29*, 841–859.

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*, 1112–1126.

Dicke, T., Parker, P. D., Holzberger, D., Kunina-Habenicht, O., Kunter, M., & Leutner, D. (2015). Beginning teachers' efficacy and emotional exhaustion: Latent changes, reciprocity, and the influence of professional knowledge. *Contemporary Educational Psychology, 41*, 62–72.

Duncan, G. J., Magnuson, K. A., & Ludwig, J. (2004). The endogeneity problem in developmental studies. *Research in Human Development, 1*, 59–80.

Dustmann, C., Puhani, P. A., & Schönberg, U. (2017). The long-term effects of early track choice. *The Economic Journal, 127*(603), 1348–1380.

De Fraine, B., Van Damme, J., & Onghena, P. (2007). A longitudinal analysis of gender differences in academic self-concept and language achievement: A multivariate multilevel latent growth approach. *Contemporary Educational Psychology, 32*, 132–150.
Ebbes, P., Wedel, M., Böckenholt, U., & Steerneman, T. (2005). Solving and testing for regressor-error (in) dependence when no instrumental variables are available: With new evidence for the effect of education on income. *Quantitative Marketing and Economics, 3*, 365–392.
Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford Press.
Erikson, R., Goldthorpe, J. H., & Portocarero, L. (1979). Intergenerational class mobility in three Western European Societies: England, France, and Sweden. *British Journal of Sociology, 30*, 341–415.
Evertson, C. M., & Weinstein, C. S. (2006). Handbook of Classroom Management: Research. Practice and Contemporary Issues. Mahwah NJ.
Van Ewijk, R., & Sleegers, P. (2010). The effect of peer socioeconomic status on student achievement: A meta-analysis. *Educational Research Review, 5*, 134–150.
Frenzel, A. C., Becker-Kurz, B., Pekrun, R., Goetz, T., & Lüdtke, O. (2018). Emotion transmission in the classroom revisited: A reciprocal effects model of teacher and student enjoyment. *Journal of Educational Psychology, 110*(5), 628–639.
Frenzel, A. C., Goetz, T., Lüdtke, O., Pekrun, R., & Sutton, R. E. (2009). Emotional transmission in the classroom: Exploring the relationship between teacher and student enjoyment. *Journal of Educational Psychology, 101*(3), 705.
Friedrich, A., Flunger, B., Nagengast, B., Jonkmann, K., & Trautwein, U. (2015). Pygmalion effects in the classroom: Teacher expectancy effects on students' math achievement. *Contemporary Educational Psychology, 41*, 1–12.
Gnas, J., Mack, E., & Preckel, F. (2022). When classmates influence teacher judgment accuracy of students' cognitive ability: Studying frame-of-reference effects in primary school. *Contemporary Educational Psychology, 69*, Article 102070.
Guldemond, H., & Bosker, R. J. (2009). School effects on students' progress – a dynamic perspective. *School Effectiveness and School Improvement, 20*(2), 255–268.
Guill, K., Lüdtke, O., & Köller, O. (2017). Academic tracking is related to gains in students' intelligence over four years: Evidence from a propensity score matching study. *Learning and Instruction, 47*, 43–52.
Guo, J., Marsh, H. W., Morin, A. J., Parker, P. D., & Kaur, G. (2015). Directionality of the associations of high school expectancy-value, aspirations, and attainment: A longitudinal study. *American Educational Research Journal, 52*, 371–402.
Guo, J., Marsh, H. W., Parker, P. D., & Dicke, T. (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.
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.
Hanushek, E. A., & Wößmann, L. (2006). Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries. *The Economic Journal, 116*(510), C63–C76.
Harker, R., & Tymms, P. (2004). The effect of student composition on school outcomes. *School Effectiveness and Improvement, 15*, 177–199.
Hattie, J. A. (2002). Classroom composition and peer effects. *International Journal of Educational Research, 37*(5), 449–481.
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.
Hörsch, H., Schumacher, J., Möller, J., & Wolff, F. (2025). Boosting students' academic self-concepts: A self-concept intervention about comparison processes. *Journal of Educational Psychology. Advance online publication.*
Jerrim, J., Parker, P. D., Katyn Chmielewski, A., & Anders, J. (2015). Private schooling, educational transitions, and early labour market outcomes: Evidence from three anglophone countries. *European Sociological Review, 32*, 280–294.
Keller, M. M., Hoy, A. W., Goetz, T., & Frenzel, A. C. (2016). Teacher enthusiasm: Reviewing and redefining a complex construct. *Educational Psychology Review, 28*(4), 743–769.
Keller, M. M., Neumann, K., & Fischer, H. E. (2013). Teacher enthusiasm and student achievement. In J. Hattie, & E. M. Andermann (Eds.), *International guide to student achievement* (pp. 247–250). New York: Routledge.
von Keyserlingk, L., Becker, M., & Jansen, M. (2019). Academic self-concept during the transition to upper secondary school. *Contemporary Educational Psychology, 56*, 152–160.
Korpershoek, H., Harms, T., de Boer, H., van Kuijk, M., & Doolaard, S. (2016). A meta-analysis of the effects of classroom management strategies and classroom management programs on students' academic, behavioral, emotional, and motivational outcomes. *Review of Educational Research, 86*(3), 643–680.
Kunter, M., Baumert, J., & Köller, O. (2007). Effective classroom management and the development of subject-related interest. *Learning and Instruction, 17*(5), 494–509.
Kunter, M., Klusmann, U., Baumert, J., Richter, D., Voss, T., & Hachfeld, A. (2013). Professional competence of teachers: Effects on instructional quality and student development. *Journal of Educational Psychology, 105*(3), 805–820.
Kunter, M., Tsai, Y. M., Klusmann, U., Brunner, M., Krauss, S., & Baumert, J. (2008). Students' and mathematics teachers' perceptions of teacher enthusiasm and instruction. *Learning and Instruction, 18*, 468–482.
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.
Lüdtke, O., Trautwein, U., Kunter, M., & Baumert, J. (2006). Reliability and agreement of student ratings of the classroom environment: A reanalysis of TIMSS data. *Learning Environments Research, 9*(3), 215–230.
Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies, 60*(3), 531–542.
Marks, G. N. (2010). What aspects of schooling are important? School effects on tertiary entrance performance. *School Effectiveness and School Improvement, 21*(3), 267–287.
Marks, G. N. (2015). Are school-SES effects statistical artefacts? evidence from longitudinal population data. *Oxford Review of Education, 41*(1), 122–144.
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.
Marsh, H. W., & Bailey, M. (1991). Multidimensional student evaluations of teaching effectiveness: A profile analysis. *Reports-Research, 143*, 800.
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*, 320–341.
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.
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.
Marsh, H. W., Pekrun, R., Dicke, T., Guo, J., Parker, P. D., & Basarkod, G. (2023). Disentangling the long-term compositional effects of school-average achievement and SES: A substantive-methodological synergy. *Educational Psychology Review, 35*(3), 70.
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. *Dev Psychol., 54*(2), 263–280. https://doi.org/10.1037/dev0000393
Marsh, H. W., Pekrun, R., Parker, P. D., Murayama, K., Guo, J., Dicke, T., & Arens, A. K. (2019). The murky distinction between self-concept and self-efficacy: Beware of lurking jingle-jangle fallacies. *Journal of Educational Psychology, 111*(2), 331–353.
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.
Marsh, H. W., Trautwein, U., Lüdtke, O., Baumert, J., & Köller, O. (2007). The big-fish-little-pond effect: Persistent negative effects of selective high schools on self-concept after graduation. *American Educational Research Journal, 44*(3), 631–669.
Matthewes, S. H. (2021). Better together? Heterogeneous effects of tracking on student achievement. *The Economic Journal, 131*(635), 1269–1307.
Muijs, D., & Reynolds, D. (2002). Teachers' beliefs and behaviors: What really matters? *The Journal of Classroom Interaction*, 3–15.
Muthén, B. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28*, 338–354.
Muthén, L. K., & Muthén, B. O. (1998–2010). *Mplus users guide*. Los Angeles, CA: Muthén & Muthén.
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.
OECD (2013), "Equity in Outcomes", in PISA 2012 Results: Excellence through Equity (Volume II): Giving Every Student the Chance to Succeed, OECD Publishing, Paris.
Parker, P. D., Dicke, T., Guo, J., Basarkod, G., & Marsh, H. W. (2021). Ability stratification predicts the size of the big-fish-little-pond effect. *Educational Researcher, 50*, 334–344.
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.
Parker, P. D., Marsh, H. W., Guo, J., Anders, J., Shure, N., & Dicke, T. (2017). An information distortion model of social class differences in math self-concept, intrinsic value, and utility value. *Journal of Educational Psychology, 110*, 445–463.
Parker, P. D., Marsh, H. W., Jerrim, J. P., Guo, J., & Dicke, T. (2018). Inequity and excellence in academic performance: Evidence from 27 countries. *American Educational Research Journal, 55*, 836–858.
Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys, 3*, 96–146.
Pekrun, R., Marsh, H. W., Suessenbach, F., Frenzel, A. C., & Goetz, T. (2023). School grades and students' emotions: Longitudinal models of within-person reciprocal effects. *Learning and Instruction, 83*, Article 101626.
Pekrun, R., Murayama, K., Marsh, H. W., Goetz, T., & Frenzel, A. C. (2019). Happy fish in little ponds: Testing a reference group model of achievement and emotion. *Journal of Personality and Social Psychology, 117*(1), 166–185.
Pekrun, R., vom Hofe, R., Blum, W., Frenzel, A. C., Goetz, T., & Wartha, S. (2007). Development of mathematical competencies in adolescence: The PALMA longitudinal study. In M. Prenzel (Ed.), *Studies on the educational quality of schools* (pp. 17–37). Waxmann.
Pokropek, A. (2015). Phantom effects in multilevel compositional analysis problems and solutions. *Sociological Methods & Research, 44*, 677–705.
Praetorius, A. K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: The German framework of three basic dimensions. *ZDM–Mathematics Education, 50*(3), 407–426.
Retelsdorf, J., Becker, M., Köller, O., & Möller, J. (2012). Reading development in a tracked school system: A longitudinal study over 3 years using propensity score matching. *British Journal of Educational Psychology, 82*(4), 647–671.
Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*, 417–458.

Sakaki, M., Murayama, K., Frenzel, A. C., Goetz, T., Marsh, H. W., Lichtenfeld, S., Wünsch, M., & Pekrun, R. (2025). Parents' academic expectations and aspirations predict students' achievement emotions. *Journal of Educational Psychology. Advance online publication.*

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.

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*, 143–146.

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research, 77*, 454–499.

Snijders, T., & Bosker, R. (1999). Multilevel modeling: An introduction to basic and advanced multilevel modeling.

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*, 191–207.

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.

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*, Article 101399.

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.

Vogl, K., & Preckel, F. (2014). Full-time ability grouping of gifted students: Impacts on social self-concept and school-related attitudes. *Gifted Child Quarterly, 58*(1), 51–68.

vom Hofe, R., Kleine, M., Blum, W., & Pekrun, R. (2005). Zur Entwicklung mathematischer Grundbildung in der Sekundarstufe I-theoretische, empirische und diagnostische Aspekte. In *Diagnostik von Mathematikleistungen-Jahrbuch der pädagogisch-psychologischen Diagnostik* (Vol. 4).

Wang, M. C., Haertel, G. D., & Walberg, H. J. (1993). Toward a knowledge base for school learning. *Review of Educational research, 63*, 249–294.

Wooldridge, J. M. (2015). *Introductory econometrics: A modern approach*. Nelson Education.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER conquest version 2.0*. Victoria, Australia: ACER Press, Australian Council for Educational Research.

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

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.