UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

How different levels of conceptualization and measurement affect the relationship between teacher self-efficacy and students' academic achievement

Zee, M.; Koomen, H.M.Y.; de Jong, P.F.

DOI
10.1016/j.cedpsych.2018.09.006

Publication date
2018

Document Version
Final published version

Published in
Contemporary Educational Psychology

Link to publication

Citation for published version (APA):
Zee, M., Koomen, H. M. Y., & de Jong, P. F. (2018). How different levels of conceptualization and measurement affect the relationship between teacher self-efficacy and students' academic achievement. *Contemporary Educational Psychology*, *55*, 189-200. https://doi.org/10.1016/j.cedpsych.2018.09.006

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:11 Nov 2022

Contemporary Educational Psychology 55 (2018) 189–200

Contents lists available at ScienceDirect

Contemporary Educational Psychology

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

How different levels of conceptualization and measurement affect the relationship between teacher self-efficacy and students' academic achievement☆

Marjolein Zee*, Helma M.Y. Koomen, Peter F. de Jong

Research Institute of Child Development and Education, University of Amsterdam, the Netherlands

ARTICLE INFO

Keywords:
Teacher self-efficacy
Reading comprehension
Mathematics achievement
Elementary school
Doubly latent multilevel structural equation modeling

ABSTRACT

Despite the common idea that teachers' self-efficacy (TSE) is associated with achievement, research findings in this area are ambiguous at best. In the current study, we took a multilevel perspective on the relationship between TSE and students' academic achievement and evaluated how different levels of conceptualization and measurement of TSE may affect this association. General and student-specific TSE scales and standardized achievement tests were administered among a sample of 360 fourth-to-sixth grade students and 49 teachers from 19 regular elementary schools across the Netherlands. Doubly latent multilevel structural equation modeling was used to test for direct relationships. Results indicated that student-level TSE was positively associated, and classroom-level TSE negatively associated with reading and math achievement. Teachers' aggregated student-specific self-efficacy was only associated with average classroom achievement in mathematics. These results illustrate how further specification of TSE scales and addressing the appropriate level of analysis may help to better explain variation in student academic outcomes and teacher self-efficacy.

1. Introduction

Among the many beliefs teachers might hold, few are as important for their behaviors and actions in class as their sense of self-efficacy (TSE). Various theoretical sources have asserted that TSE, or teachers' self-referent capability judgments, are likely to determine the type of classroom activities teachers choose to get into, the effort they expend in such activities, and the extent to which they persevere in difficult classroom environments (Bandura, 1977; Tschannen-Moran & Woolfolk Hoy, 2001; Tschannen-Moran, Woolfolk Hoy, & Hoy, 1998). Moreover, teachers with higher self-reported levels of efficacy have been shown empirically to use more diverse instructional strategies, to differentiate more frequently, and to be better attuned to students' signals, needs, and expectations than teachers who lack such beliefs (e.g., Hardré & Sullivan, 2008; Martin, Sass, & Schmitt, 2012; Nie, Tan, Liau, Lau, & Chua, 2013; Thoonen, Sleegers, Oort, Peetsma, & Geijsel, 2011). Perhaps not surprisingly, these findings have long nourished the conviction that TSE may be a powerful predictor of students' academic achievement as well (cf. Klassen, Tze, Betts, & Gordon, 2011; Tschannen-Moran & Woolfolk Hoy, 2001).

The results of empirical studies on the association of TSE with students' academic achievement seem, however, to be equivocal at best (Klassen & Tze, 2014; Zee & Koomen, 2016). In a recent meta-analysis of 31 empirical studies on TSE (Klassen & Tze, 2014), for instance, the mean correlation between teachers' general self-efficacy and students' overall achievement was only 0.08. Although this small correlation was statistically significant, the effect sizes of individual studies suggest that the association between TSE and students' academic performance may vary considerably. Whereas the highest reported correlation of general TSE and student achievement in Klassen and Tze's study was 0.70 (see Ross, 1992), other correlations were far weaker and demonstrated extensive variation, ranging from −0.25 (see Eberle, 2011) to 0.35 (see Heneman, Kimball, & Milanowksi, 2006).

Prior empirical research on students' achievement in specific subject areas seems to substantiate these meta-analytic findings. Some of these studies noted positive relations between general TSE and students' performance in reading, science, and mathematics (e.g., Allinder, 1995; Cantrell, Almasi, Carter, & Rintamaa, 2013; Guo, McDonald Connor, Yang, Roehring, & Morrison, 2012; Lumpe, Czerniak, Haney, & Beltyukova, 2012; Throndsen & Turmo, 2013). In several other studies,

☆ This research was supported by grant 411-12-036 from the Netherlands Organization for Scientific Research.

* Corresponding author at: Research Institute of Child Development and Education, University of Amsterdam, P.O. Box 15776, NL-1001 NG Amsterdam, the Netherlands.

E-mail address: M.Zee@uva.nl (M. Zee).

https://doi.org/10.1016/j.cedpsych.2018.09.006

Available online 25 September 2018
0361-476X/ © 2018 Elsevier Inc. All rights reserved.

however, the hypothesis that TSE may contribute to students' achievement in these subjects could not be supported (Heneman et al., 2006; Reyes, Brackett, Rivers, White, & Salovey, 2012; Tournaki & Podell, 2005).

At present, it is still unclear why the relationship between TSE and students' academic achievement is not nearly as substantial and consistent as theorists have thus far assumed. Perhaps one aspect of the complexity is that the association of TSE with academic achievement may be reciprocal in nature and affected by multiple factors that mediate or moderate this relationship (Zee & Koomen, 2016). Following Bandura's (1997) theorizing, for instance, TSE is likely to be part of a complex system of triadic reciprocal causality, in which environmental forces, personal factors, and behaviors influence one another bi-directionally. Partial support for this contention has been provided by Guo et al. (2012), whose results suggest that teachers' support for learning might mediate the association between TSE and fifth-graders' literacy outcomes. Another, but as yet unexplored reason for these inconsistent effects may be the complex, multilevel nature of the relationship between TSE and academic performance (e.g., Bandura, 2012). Based on Marsh' (1986) internal/external frame of reference (I/E) model, it can be assumed that the way TSE and achievement are conceptualized and measured at various levels may shed additional light on the somewhat equivocal results found in prior research. In the current study, therefore, we take a multilevel perspective on the relationship between TSE and students' academic achievement to evaluate how different levels of conceptualization and measurement of TSE may affect this association (e.g., Klein & Kozlowski, 2000; Marsh et al., 2009, 2012; Piantadosi, Byar, & Green, 1988).

1.1. The multilevel nature of the relationship between TSE and achievement

A common aim of research in the field of educational psychology is to explore whether characteristics at the group level (e.g., schools or classrooms) affect outcomes above and beyond what can be explained by attributes of individuals (e.g., students or teachers). Although this aim seems relatively straightforward, its elaboration may be notoriously complex as group-level characteristics can either be based on *true constructs* (e.g., teachers' gender) or *contextual constructs,* which reflect *aggregated responses* from students within a classroom (e.g., class-average ratings of teacher support; Marsh et al., 2009, 2012). This is also true for research on the link between TSE and academic achievement. More specifically, teachers' general self-efficacy seems to be a *true* L2 construct, reflecting their beliefs in their abilities to perform daily school activities in a particular classroom (e.g., Bandura, 1997; Tschannen-Moran & Woolfolk Hoy, 2001). Academic achievement, however, is usually an aggregate of the academic performance of students within a particular classroom. This indicates that all individual achievement scores at the student level (L1) are summed to form a new *contextual* construct, measured at L2 (Lüdtke et al., 2008; Marsh et al., 2012). From a multilevel perspective, then, the relationship between TSE and achievement pertains to a relationship at L2, the classroom.

Associations between variables that reflect a combination of true L2 constructs and aggregated responses are, perhaps not surprisingly, best measured within a multilevel framework (Raudenbush & Bryk, 2002). Yet in some empirical studies on TSE, the multilevel nature of the relation between TSE and academic achievement seems to be largely overlooked. For instance, Lumpe et al. (2012) and Mojavezi and Poodineh Tamiz (2012) treated TSE as a true L2 construct and academic achievement as a L1 variable in single, i.e. student-level, analyses. In such analyses, the observed relationship between TSE and achievement is still based on mean classroom performance, yet the standard errors in this statistical approach tend to be underestimated (e.g., Snijders & Bosker, 1999). Moreover, the primary interpretations that arise from these single-level models may be largely invalid, given that cross-level effects tend to depend on the amount of L2 variance (cf. Hedges & Hedberg, 2007; Piantadosi et al., 1988; Snijders & Bosker, 1999). For instance, there is some evidence that classroom contextual factors, including achievement level, class size, and student composition, may affect both students' performance and teachers' level of self-efficacy (e.g., Raudenbush, Rowan, & Cheong, 1992; Sanders, Wright, & Horn, 1997; Tsouloupas, Carson, Matthews, Grawitch, & Barber, 2010). Hence, differences among results on the association between TSE and achievement may be partly due to whether this relationship has been evaluated in a single-level framework considering only the student level, or in frameworks that appropriately address the multilevel nature of the data (Marsh et al., 2012).

Another issue concerns the extent to which TSE truly reflects an L2 construct and is a proper reflection of a teacher's sense of self-efficacy toward the particular class involved. Within the context of Bandura's (1997) social-cognitive theory, teachers' feelings of self-efficacy have typically been conceptualized as their beliefs about their existing abilities, resources, and opportunities to perform daily school activities in a *particular classroom* (e.g., Tschannen-Moran & Woolfolk Hoy, 2001). At first glance, TSE thus seems to be a true L2 construct. Yet unlike other true L2 constructs, such as teachers' gender, classroom size, or grade level, this interpretation is not entirely straightforward. Following Bandura's (1977, 1997) theorizing, it seems highly plausible that TSE is conditional upon various psychological processes, such as enactive attainments in particular classes (*mastery experiences*), referential comparisons (*vicarious experiences*), or external feedback from students and colleagues (*verbal persuasion*). Integrating Bandura's notions with Marsh' (1986) internal/external frame of reference (I/E) model, such processes can be further classified into *internal* and *external* comparisons, both of which may ultimately affect the nature of TSE and its associations with achievement in the classroom.

Generally, external, or normative comparisons reflect a psychological process in which teachers use external norm criteria to compare their self-efficacy toward a particular classroom with their self-efficacy for other classrooms they have taught. These criteria include, among others, their enactive attainments, the levels of stress and fatigue they experience in a particular classroom, or the degree of challenge students in class provide (cf. Bandura, 1997; Marsh, 1986). Conceivably, teachers may use such environmental cues as a basis for determining their level of self-efficacy across classrooms. When teachers compare their self-efficacy and performances in a particular classroom with relatively easier classrooms they have taught, they may, for instance, lower their self-efficacy, thereby possibly negatively affecting their students' academic performance towards this class as well. Yet, when teachers' self-efficacy for a particular class is higher on the basis of the external norm criteria they have in mind, their impact on student performance is most likely to be positive. As such, TSE seems, at least in part, to reflect a combination of teachers' beliefs in their capability to teach a particular class, and the external frames of reference teachers use to evaluate these beliefs (Bandura, 1997). This may potentially cloud the predictability of classroom-average achievement, which usually is solely based on aggregated test scores from students within a classroom.

The internal, or dimensional process may refer to comparisons teachers make across students in evaluating their self-efficacy (Marsh, 1986; Zee, Koomen, Jellesma, Geerlings, & de Jong, 2016). There is an emerging body of evidence supporting the social-cognitive view that teachers' self-efficacy may vary as a function of individual students in the classroom (e.g., Geerlings, Thijs, & Verkuyten, 2017; Zee et al., 2016; Zee, de Jong, & Koomen, 2016, 2018). Such multidimensional perceptions may lead to *internal comparisons* that affect the role of TSE in students' achievement. If, for instance, teachers' self-efficacy for one particular student in class is higher than for another, this student's academic achievement might also improve more than the achievement of the other. This so-called frame-of-reference effect may explain why some students may achieve less than other students in the classroom, due to teachers' less favorable student-specific self-efficacy beliefs (cf. Marsh, 1991).

The idea of internal/external comparisons and corresponding frame-of-reference effects may have implications for the measurement and analysis of TSE in relation to achievement. First, the items teachers have to respond to in many self-efficacy questionnaires may be open to interpretation as they usually do not tap external or internal norm criteria. Consider, for example, an item of Tschannen-Moran and Woolfolk Hoy's (2001) well-validated and often used Teacher Sense of Efficacy Scale (TSES): "How well can you respond to difficult questions from your students?". Teachers' answer to this item probably depends on the specific classroom circumstances under which they have to respond to current students' questions, their experiences with other classes they have taught, as well as on the very students who ask those questions. Unfortunately, we generally do not know which specific students or situations teachers have in mind when responding to self-efficacy items (cf. Bandura, 2006; Wheatley, 2005; Zee et al., 2016). Teachers may thus differ in how they interpret the same self-efficacy item, which would result in increases in measurement error and corresponding decreases in validity (Bing, Whanger, Davison, & VanHook, 2004). This obviously affects the link between TSE and achievement as well.

Second, without common frames of reference (i.e., comparisons) in self-efficacy items, it is unclear to what extent TSE-scores are a proper reflection of teachers' feelings toward the entire classroom if, at least with respect to some items, some of the students may be more influential than others. Specifically, if answers to some TSE questions are based on the characteristics of some individual students, then general TSE might be (partly) regarded as an implicit aggregate of teachers' sense of self-efficacy toward *particular students* and thus as a contextual L2 construct instead of a true L2 construct. Thus, unlike classroom-level achievement, it is uncertain how teachers' responses are aggregated to the L2 level, and which students teachers have in mind when responding to items about their self-efficacy toward a particular classroom (cf. Bandura, 2006; Wheatley, 2005; Zee et al., 2016).

1.2. A full multilevel approach to the relationship between TSE and achievement

Inspired by the theoretical and empirical work of Bandura (1997) and Marsh (Marsh, 1986; Marsh et al., 2009, 2012), in this paper we explicitly conceive teachers' sense of self-efficacy as a contextual L2 construct, rather than a true L2 construct. Our approach entails that teachers report on their sense of self-efficacy toward each *individual student* in their classroom, denoted by Zee et al. (2016) as student-specific self-efficacy. In doing so, we depart from previous studies on TSE in three important ways. First, by measuring both TSE and students' academic achievement at L1, TSE and achievement at L2 can be aggregated over the same students. These aggregated scores are likely to reflect teachers' feelings of self-efficacy toward their students in a particular classroom. Second, by contextualizing items to individual students, we may provide clearer and more standardized item interpretations that may help teachers respond more accurately to self-efficacy items (see Bing et al., 2004, Marsh, 1986, and Zee et al., 2016, for similar arguments). Third, unlike general (classroom-level) TSE, which only pertains to L2, conceiving TSE as a contextual L2-construct provides the opportunity to investigate the association between TSE and students' academic achievement both at the classroom and the student level (Klein & Kozlowski, 2000; Marsh et al., 2012). To our knowledge, the association between student-specific TSE and students' academic achievement at L1 has not been examined before.

1.3. Present study

In this study, we aim to examine the association between TSE and students' achievement by using a multilevel approach. To examine how different levels of conceptualization and measurement of TSE (true or contextual) may affect the relation of TSE with students' achievement, we incorporated the oft-used original TSES (Tschannen-Moran & Woolfolk Hoy, 2001) to evaluate teachers' *general* self-efficacy, and the Student-Specific TSES (Zee et al., 2016) to measure teachers' self-efficacy toward *individual students*. Both measures were evaluated in relation to students' performance in the main subject areas of reading comprehension and mathematics.

We used multilevel structural equation models to simultaneously examine relationships at the classroom and student level (e.g., Lüdtke et al., 2008; Marsh et al., 2012; Morin, Marsh, Nagengast, & Scalas, 2014). At the classroom level (L2), we examined the relationships of teachers' general self-efficacy and teachers' aggregated student-specific self-efficacy with aggregate achievement in reading and mathematics. Additionally, at the student level (L1), we evaluated the association of student-specific TSE with individual students' reading comprehension and math achievement. This allowed us to compare the relationship between TSE and students' achievement at L1 and L2. Thereby, this study may provide gentle impetus for longitudinal studies investigating causal links between TSE and achievement.

2. Method

2.1. Participants and procedure

This study included data from 49 teachers and 360 students in fourth-to-sixth grade classrooms from 19 regular elementary schools located in urban and rural areas in the Netherlands. This investigation was approved by the institutional Ethics Review Board (project no. 2013-CDE-3188) and consent forms from teachers and parents were obtained prior to participation. From all signed consents, four boys and four girls were randomly selected from each teacher's classroom[1] about whom teachers completed questionnaires. Due to illness or time constraints, eight of 49 teachers completed questionnaires about fewer children (range = 2–7 students). The decision to randomly select eight students per classroom was based on guidelines from Snijders and Bosker (1999), who have indicated that relatively high intra-class correlations may decrease the benefits of including whole classes in the sample. Moreover, including more students per class would make the data collection overly burdensome for teachers and would compromise their willingness to participate.

During a planned school visit in the second half of the schoolyear, teachers were asked to fill out the general TSE measure at the classroom level as well as several questions about background demographics. To avoid common method variance, teachers completed the student-specific teacher self-efficacy questionnaire for the eight randomly selected students through an online survey. The anonymous link to this survey was e-mailed directly after the school visit. Teachers were asked to complete the student-specific self-efficacy questionnaire within two weeks. Students' reading comprehension and mathematics achievement were assessed during the same period. Their standardized test scores were obtained through teachers' classroom database, several weeks after the planned school visits.

Among the total sample of students were 180 boys (50%) and 180 girls (50%) from grades 4 (n = 119), 5 (n = 139), and 6 (n = 102), respectively. These children ranged from 8.4 to 13.0 years of age (M = 10.7, SD = 0.95) and most of them had a Dutch nationality (90.3%). Based on teacher reports of parents' employment status and

[1] Comparisons of the selected sample with the total sample containing whole classrooms revealed no statistically significant mean differences in terms of students' ethnicity ($t(1860) = 1.36$, $p = .174$) and age ($t(1885) = -0.60$, $p = .546$). Moreover, there were no significant group differences across the samples in terms of reading comprehension ($t(1492) = 1.00$, $p = .319$), and math achievement ($t(1360) = 1.89$, $p = .058$). Based on these results, the selected sample can be considered representative for the total sample with respect to the study's main variables.

educational level, 23 students (6.4%) were considered to have a low SES, 191 students (53.1%) a medium SES, and 133 students (36.9%) a high SES. For 13 students, information about SES was not available. The teacher sample consisted of 36 females (74.2%). Teachers had a mean age of 41.2 (SD = 12.2, range = 23.3–63.3) and their years of professional teaching experience ranged from 1.5 to 44 .0 years (M = 16.3, SD = 12.0). Of note, these classroom teachers taught a range of subjects, including reading and math, to all students in the classroom, which is common practice in the Netherlands.

2.2. Instruments

2.2.1. General TSE

Teachers' perceptions of their general self-efficacy were estimated using the short form of the Teachers' Sense of Efficacy Scale (TSES; Tschannen-Moran & Woolfolk Hoy, 2001). This instrument evaluates teachers' perceived capability across three dimensions: Instructional Strategies (IS; 4 items), Classroom Management (CM; 4 items), and Student Engagement (SE; 4 items). Yet, to allow for accurate comparisons between teachers' general and student-specific self-efficacy, we only incorporated the parallel self-efficacy domains of IS (4 items) and SE (4 items), which are included in both the general and student-specific TSES (see Appendix A).

Examples of items of the IS and SE domains are "To what extent can you provide an alternative explanation or example when students are confused?" and "How much can you do to help your students value learning?", respectively. Teachers were asked to respond to all items in reference to their current classroom on a 7-point Likert scale, ranging from 1 (*not at all*) to 7 (*a great deal*). The factorial validity of the dimensions of the short TSES has been shown to be adequate and evidence of measurement invariance has been found across grades and countries (Klassen et al., 2009; Tschannen-Moran & Woolfolk Hoy, 2001). In the present study, Cronbach's alpha was 0.71 for IS and 0.75 for SE.

2.2.2. Student-specific TSE

Teachers rated their student-specific self-efficacy beliefs using the Student-Specific Teacher Self-Efficacy Scale (Zee et al., 2016). This instrument reflects TSE in relation to individual students across four teaching domains, including Instructional Strategies (6 items), Student Engagement (6 items), Behavior Management (5 items), and Emotional Support (7 items). In this study, we selected the four Instructional Strategies and four Student Engagement items that parallel those of the general TSES (see Appendix A). Example items for each domain are "To what extent can you provide an alternative explanation or example when this student is confused?" and "To what extent can you help this student to value learning?", respectively. All items were rated on a 7-point Likert scale, ranging from 1 (*not at all*) to 7 (*a great deal*). Preliminary support for the construct validity of the Student-Specific TSES has been provided by Zee et al. (2016). The internal consistency of the IS-scale (α = 0.87) and SE-scale (α = 0.88) was satisfactory.

2.2.3. Students' reading comprehension and math achievement

We used the official results of national tests of reading comprehension and mathematics achievement for first-to-sixth graders (LOVS; Leerling- en Onderwijs Volgsysteem, [System for the Longitudinal Assessment of School Achievement]), developed by the Dutch national institute for assessment in education, CITO. Both instruments are nationally normed and well-validated achievement tests, developed to screen and determine students' current level of reading comprehension and mathematics achievement (Janssen, Verhelst, Engelen, & Scheltens, 2010; Weekers, Groenen, Kleintjes, & Feenstra, 2011). Reliability coefficients of both tests have been shown to be adequate, ranging from 0.91 to 0.97 for math (Janssen et al., 2010), and from 0.87 to 0.89 for reading comprehension (Weekers et al., 2011). We used standardized ability scores of the reading comprehension and math tests, which are based on item response theory and take the number and complexity of items of the tests into account. Scores were obtained through teachers' classroom database.

2.3. Data analysis

Using M*plus* version 7.11 (Muthén & Muthén, 1998–2012), we applied doubly latent multilevel structural equation models (ML-SEMs; Lüdtke, Marsh, Robitzsch, & Trautwein, 2011; Marsh et al., 2009, 2012) to examine the associations of classroom-level and student-specific TSE with reading comprehension and math achievement. ML-SEM takes the clustering of students within the teacher into account by partitioning the variation in students' achievement between and within teachers (Snijders & Bosker, 1999). Thereby, it allows for the calculation of unbiased estimates of the standard errors associated with the regression coefficients and for the inclusion of between-teacher characteristics (i.e., aggregated and general TSE) in models of individual student outcomes (i.e., students' achievement). In addition, doubly latent refers to the ability of these models to control for measurement and sampling error. Similar to confirmatory factor analysis, observed variables are specified as indicators of latent variables. Moreover, the aggregate over the eight randomly selected L1 student scores per classroom are considered as an estimate of the latent score of the construct at L2. Put differently, the student scores are conceived as items in a scale and, as in classical test theory, the mean score, in this case the average, or aggregated scores of the eight randomly selected students within each participating classroom, can be considered as an estimate of the score on the underlying latent trait.

2.3.1. Modeling procedure

Data were analyzed in three steps. First, to justify the use of doubly latent ML-SEM, we evaluated the variability at L2 and calculated corresponding intraclass correlation coefficients (ICC1). Also, we evaluated the reliability of the aggregated scores for the two achievement measures as well as student-specific TSE at L2 (ICC2; Marsh et al., 2009, 2012; Raudenbush & Bryk, 2002). In our study, these aggregated scores of the eight randomly selected students within each classroom can be considered as an estimate of the classroom mean. In the second step, we performed doubly latent confirmatory factor analysis to find evidence for the hypothesized factor structure of the measures at L1 and L2. In the last step, we specified a full doubly latent ML-SEM with both general and aggregated TSE at the classroom level (L2) and student-specific TSE at the student level (L1). Students' gender and age were included as student-level covariates, and teachers' gender and years of teaching experience as teacher-level covariates. All variables that were included at both L1 and L2 were group-mean centered. Covariates specified at L1 were grand-mean centered (Marsh et al., 2012).

2.3.2. Model goodness-of-fit

Maximum likelihood estimation with robust standard errors and a scaled test statistic (MLR) was chosen as the estimation method (Muthén & Muthén, 1998–2012). Additionally, missing data (< 10%) were treated using full information maximum likelihood estimation (FIML). Under assumptions of data missing-at-random (MAR), the combination of these estimation methods has been demonstrated to lead to unbiased parameter estimates (e.g., Enders & Bandalos, 2001; Shin, Davidson, & Long, 2009). Overall goodness-of-fit of the models was evaluated by the mean-adjusted χ^2-test, with non-significant chi-squares indicating satisfactory fit. Approximate fit was determined using the Root Mean Square Error of Approximation (RMSEA), with values below 0.05 reflecting close fit, and below 0.08 signifying

reasonable fit (Browne & Cudeck, 1992), and the Comparative Fit Index (CFI), with values ≥0.90 indicating satisfactory fit, and values ≥0.95 indicating close fit (Bentler, 1992). The model's modification indices, residual correlations, and their associated summary statistic SRMR (Standardized Root Mean Square Residual) were used to evaluate component fit. Values ≤0.08 indicate good model fit (Kline, 2011).

3. Results

3.1. Intraclass correlations

Prior to main analyses, we inspected the intraclass correlations (ICC1), as well as the reliability of the aggregated classroom scores (ICC2; Marsh et al., 2009, 2012; Raudenbush & Bryk, 2002). The ICC1 refers to the proportion of the total variance in Reading Comprehension and Math Achievement that is accounted for by the clustering of students within classrooms. This coefficient thus reflects the agreement between students within the same classroom. In this study, the ICC1 values ranged between 0.18 (Student-Specific TSE-item 6) and 0.42 (Mathematics Achievement), suggesting that between 18% and 42% of the variance occurred between classrooms.

Additionally, the ICC2 can be used to determine the reliability of aggregated L2 measures (i.e., the averages for Student-Specific TSE, Reading Comprehension, and Math Achievement), based on the scores of the eight randomly selected students within each participating classroom (Lüdtke et al., 2008; Marsh et al., 2009). Following the guidelines of Cicchetti (1994), ICC2 values between 0.60 and 0.74 can be considered as good, and between 0.75 and 1.00 as excellent. In this study, the ICC2 values were 0.78 for the aggregated Student-Specific TSE-scores, 0.88 for aggregated Math Achievement, and 0.85 for aggregated Reading Comprehension, respectively. Together, the ICC1 and ICC2 indexes supported the excellent reliability of the aggregated variables and justified the use of doubly latent ML-SEM (Marsh et al., 2012).

Table 1
Latent factor correlations among student-level (L1) constructs.

	1.	2.	3.	4.	5.
Covariates					
1. Student Age	1.00				
2. Student Gender	−0.14 (0.04)***	1.00			
Teachers' Self-efficacy					
3. Student-Specific TSE	−0.38 (0.11)***	0.22 (0.06)***	1.00		
Students' Achievement					
4. Mathematics Achievement	−0.001 (0.14)	−0.12 (0.05)**	0.43 (0.08)***	1.00	
5. Reading Comprehension	0.05 (0.12)	0.11 (0.06)	0.32 (0.05)***	0.59 (0.05)***	1.00
Descriptive Statistics					
Mean	10.73	–	5.58	9.80	4.65
Standard Deviation	0.95	–	0.94	1.64	1.95
Range	8.40–13.00	–	2.13–7.00	3.90–13.0	0.60–10.90

Note. $N_{teachers}$ = 49; $N_{students}$ = 360. Standard errors are displayed between brackets. Means and standard deviations are based on the total variance-covariance matrix. Gender: 0 = boys, 1 = girls. TSE = teacher self-efficacy.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

Table 2
Latent factor correlations among classroom-level (L2) constructs.

	1.	2.	3.	4.	5.	6.
Covariates						
1. Tenure	1.00					
2. Teacher Gender	−0.37 (0.16)*	1.00				
Teachers' Self-efficacy						
3. General TSE	0.21 (0.14)	−0.05 (0.16)	1.00			
4. Aggregated TSE	0.40 (0.17)*	−0.24 (0.16)	0.57 (0.15)***	1.00		
Students' Achievement						
5. Mathematics Achievement	0.24 (0.13)	−0.22 (0.14)	−0.18 (0.15)	0.45 (0.17)**	1.00	
6. Reading Comprehension	0.16 (0.16)	−0.19 (0.15)	−0.24 (0.15)	0.23 (0.25)	0.72 (0.21)***	1.00
Descriptive Statistics						
Mean	16.29	–	5.38	5.58	9. 80	4.65
Standard Deviation	12.01	–	0.59	0.59	1.20	1.39
Range	1.50–44.00	–	3.88–6.38	3.59–6.97	6.69–11.63	2.28–9.04

Note. $N_{teachers}$ = 49; $N_{students}$ = 360. Standard errors are displayed between brackets. Aggregated TSE is based on aggregations of student-level (L1) student-specific teacher self-efficacy scores. Means and standard deviations are based on the total variance-covariance matrix. Gender: 0 = males, 1 = females. TSE = teacher self-efficacy.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

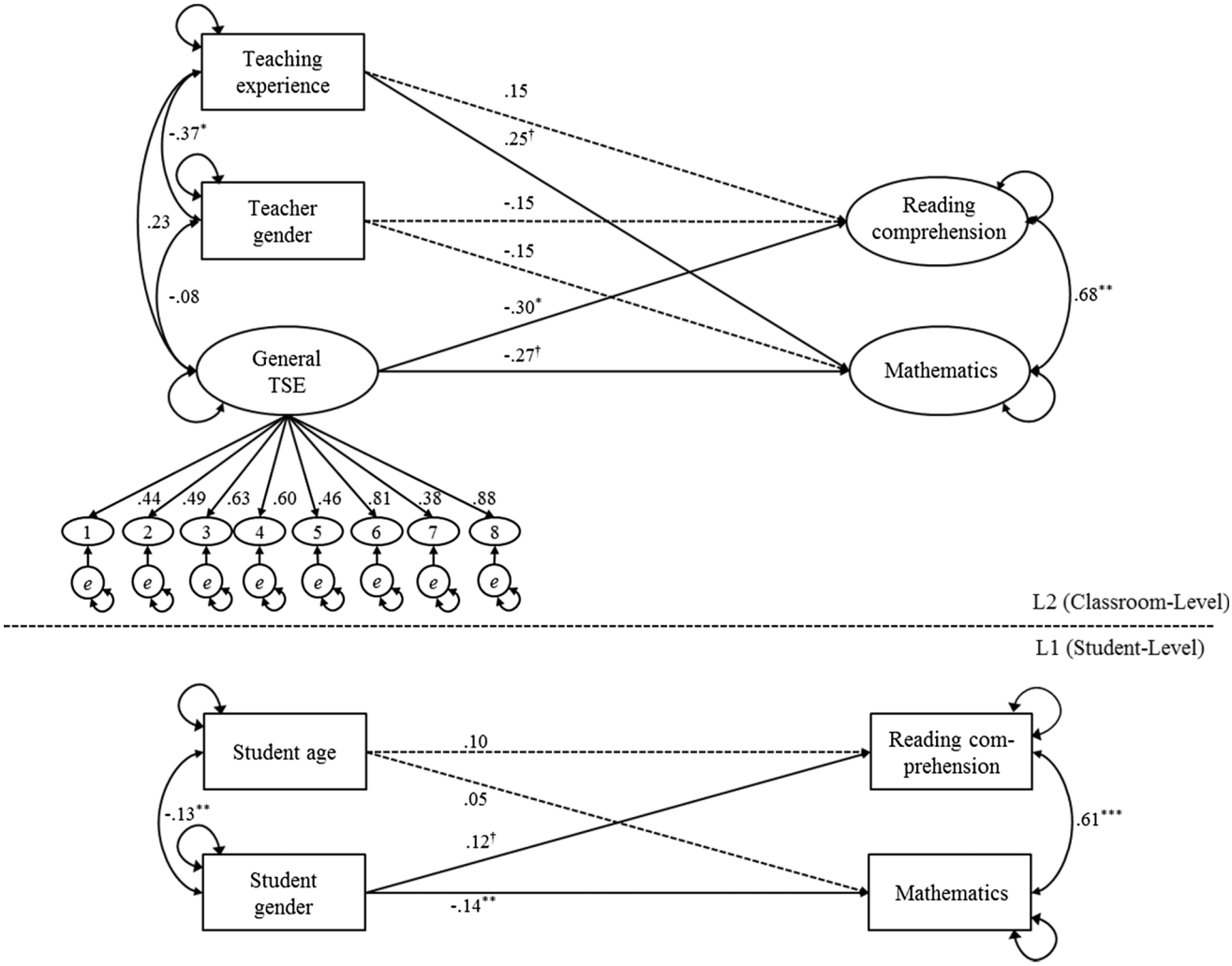

Fig. 1. Doubly latent multilevel structural equation model of general TSE. *Note.* Standardized robust maximum likelihood parameter estimates are reported. Dashed lines represent non-significant paths. Gender: 0 = boy/male, 1 = girl/female. TSE = Teacher Self-Efficacy. $^{\dagger}p < .10$, $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

3.2. Doubly latent measurement model and inter-factor correlations

Next, we evaluated the underlying measurement model of General and Aggregated TSE at L2 and of Student-Specific TSE at L1. In this model, the latent factors of IS and SE at each respective level of analysis were allowed to correlate and the factor loadings of the indicators of Aggregated and Student-Specific TSE were freely estimated across L1 and L2. The model demonstrated a reasonable fit to the data, $\chi^2(189) = 402.52$, $p < .001$, RMSEA = 0.057, CFI = 0.912, $SRMR_{within} = 0.041$, $SRMR_{between} = 0.085$. Yet, parameter estimates indicated substantial correlations between the IS and SE-factors for Student-Specific TSE at L1 ($r = 0.95$, $p < .001$), Aggregated TSE at L2 ($r = 0.97$, $p < .001$), and General TSE at L2 ($r = 0.76$, $p < .001$).

Consistent with raw correlations and prior research (Klassen et al., 2009; Zee et al., 2016, 2018), these findings suggest that the domains of IS and SE probably tap similar constructs. Therefore, we fitted a model with one Student-Specific TSE factor at L1, and one General TSE and one Aggregated TSE factor at L2. This model fitted the data well, but could be further improved by adding a residual correlation between Instructional Strategies items 3 and 4 at L1 and L2 (see Appendix A). This resulted in a satisfactory model fit, $\chi^2(205) = 370.10$, $p < .001$, RMSEA = 0.048, CFI = 0.932, $SRMR_{within} = 0.051$, $SRMR_{between} = 0.113$.

To verify that Student-Specific TSE at L1 and Aggregated TSE at L2 were measured in the same metric, we tested the invariance of their factor loadings across levels (Marsh et al., 2012). Imposing equality constraints across levels did not significantly deteriorate the model's fit, $\chi^2(212) = 391.46$, $p < .001$, RMSEA = 0.049, CFI = 0.926, $SRMR_{within} = 0.035$, $SRMR_{between} = 0.099$, suggesting measurement invariance across levels. Hence, these results suggest that the Student-Specific and Aggregated TSE constructs are similar at both levels and that contextual effects can be properly estimated (Morin et al., 2014).

In the final measurement model, the factor loadings of the items on the General TSE factor at L2 ranged from 0.34 to 0.86 and its composite reliability (see Brown, 1989) was 0.81. Standardized factor loadings of the Student-Specific TSE factor, ranging from 0.67 to 0.80, and the Aggregated TSE factor, ranging from 0.74 to 1.00, were generally higher than those of the General TSE factor. The composite reliability of both latent factors was excellent, $\rho_c = 0.91$ for Student-Specific TSE and $\rho_c = 0.97$ for Aggregated TSE, respectively. Hence, despite the relatively low cluster size, the psychometric properties of the scales in this study indicate that we used ML-SEM under very favorable conditions

(Lüdtke, Robitzsch, Trautwein, & Kunter, 2009).

Patterns of correlations among the latent factors are presented in Tables 1 and 2, respectively. At L1, teachers' Student-Specific Self-Efficacy was significantly and positively correlated with students' Reading Comprehension ($r = 0.32$, $p < .001$) and Math Achievement ($r = 0.43$, $p < .001$). The correlation between the two achievement measures ($r = 0.59$, $p < .001$) was also positive and moderate in nature. At L2, we found negative, though non-significant associations among General TSE and the two aggregated achievement scores.[2] Aggregated TSE (i.e., Student-Specific TSE aggregated to the classroom level of analysis), however, was positively associated with aggregated Math Achievement ($r = 0.45$, $p < .01$), but not related to Reading Comprehension.

It is interesting to note that General and Aggregated TSE were only moderately correlated ($r = 0.57$, $p < .001$), suggesting that these constructs at L2 may tap different aspects of the teacher self-efficacy belief system. This idea is supported by the finding that both measures have a different pattern of association with aggregated Reading Comprehension and Math Achievement. Specifically, constraining the correlations of General and Aggregated TSE with aggregated Math Achievement to be equal resulted in a statistically significant deterioration in model fit, $\Delta\chi^2(1) = 32.77$, $p < .001$, ΔCFI = 0.013. Similar differences were noted for the correlations of General and Aggregated TSE with aggregated Reading Comprehension, $\Delta\chi^2(1) = 5.93$, $p < .05$, ΔCFI = 0.002. Also, constraining the correlations of General TSE with the two achievement measures and the correlations of Aggregated TSE with the achievement measures to be equal resulted in a worse model fit, $\Delta\chi^2(3) = 16.45$, $p < .05$, ΔCFI = 0.009. Thus, even though General and Aggregated TSE are moderately correlated with each other, they appear to have different links with Reading Comprehension and Math Achievement at the classroom level.

3.3. Doubly latent multilevel structural equation model

In the next step, we fitted a full doubly latent ML-SEM with General and Aggregated TSE as latent teacher-level factors, and Student-Specific TSE as an individual student-level factor. This model had an acceptable fit to the data, $\chi^2(212) = 386.43$, $p < .001$, RMSEA = 0.049, CFI = 0.928, $SRMR_{within} = 0.035$, $SRMR_{between} = 0.099$. Path estimates at L1 indicated statistically significant positive associations among Student-Specific TSE and individual students' Reading Comprehension ($\beta = 0.38$, $p < .001$) and Math Performance ($\beta = 0.55$, $p < .001$). At L2, however, we obtained unexpectedly high and statistically significant negative associations of teachers' General Self-Efficacy with aggregated Reading Comprehension ($\beta = -0.53$, $p < .01$) and Math Performance ($\beta = -0.63$, $p < .01$). Moreover, Aggregated TSE showed a strong, positive association with Mathematics Achievement ($\beta = .78$, $p < .001$) and Reading Comprehension at the classroom level, although the latter was only marginally significant ($\beta = 0.50$, $p = .073$).

Given the far weaker inter-factor correlations in Table 2, this pattern of results seems to represent a situation in which both General and Aggregated TSE suppress part of each other's information that is irrelevant to the two achievement measures at L2. In this case of *reciprocal suppression* (Maassen & Bakker, 2001), the absolute values of the path estimates for the association of Aggregated TSE with aggregated scores for Reading Comprehension and Math Achievement are likely to increase when General TSE is entered as a second predictor. To avoid this suppression phenomenon and obtain more reliable results, we continued by fitting two separate doubly latent ML-SEMs for General TSE and Aggregated TSE, respectively.

3.3.1. Doubly latent ML-SEM for general TSE

The model with General TSE, teachers' Gender, and Teaching Experience at L2, and students' Age and Gender at L1, had a good fit to the data, $\chi^2(48) = 59.41$, $p = .125$, RMSEA = 0.026, CFI = 0.951, $SRMR_{within} = 0.008$, $SRMR_{between} = 0.093$. Considering the model's modification indices, no further model adjustments were needed to improve the model's fit. Fig. 1 displays the standardized coefficients for the final doubly latent structural model for General TSE. At L1, girls appeared to perform worse in Math than boys ($\beta = -0.14$, $p < .05$), and marginally better in Reading Comprehension than boys ($\beta = 0.12$, $p = .055$). At L2, we found a statistically significant negative association between General TSE and aggregated Reading Comprehension scores in the classroom ($\beta = -0.30$, $p < .05$). The negative association between General TSE and aggregated Math Achievement was marginally significant ($\beta = -0.27$, $p = .080$). Overall, both covariates and General TSE explained 12.4% of the variance in aggregated Reading Comprehension, and 14.9% of the variance in aggregated Math scores.

3.3.2. Doubly latent ML-SEM for aggregated and student-specific TSE

In this step, we fitted the model with Student-Specific TSE and student-level covariates at L1, Aggregated TSE and teacher-level covariates at L2, and invariant factor loadings of the TSE-indicators across levels. The fit of this final model was satisfactory, $\chi^2(101) = 206.89$, $p < .001$, RMSEA = 0.055, CFI = 0.947, $SRMR_{within} = 0.036$, $SRMR_{between} = 0.084$. The standardized coefficients are displayed in Fig. 2. As in the model with General TSE, at L1, girls appeared to perform worse in Math than boys ($\beta = -0.22$, $p < .001$), and older students performed better in Math ($\beta = 0.18$, $p < .10$) and Reading Comprehension ($\beta = 0.20$, $p < .05$) than younger students. After accounting for students' Age and Gender, Student-Specific TSE was significantly and positively associated with individual students' Reading Comprehension ($\beta = 0.38$, $p < .001$) and Math Achievement ($\beta = 0.55$, $p < .001$) at L1. After properly controlling for these associations at L1, the small to moderate association of Aggregate TSE with aggregated Reading Comprehension ($\beta = 0.19$, *ns*) was not significant, whereas the association with aggregated Math Achievement ($\beta = 0.40$, $p = .066$) was marginally significant.

We also calculated the contextual effect of teachers' (aggregated) Student-Specific Self-Efficacy beliefs on the two achievement measures at L2 by creating two additional parameters. These parameters reflect the difference between the corresponding L2 and L1 path coefficients for the associations of TSE with Reading Comprehension and Math. Following Marsh et al. (2012), these contextual effects are likely to be more reliable as they take the conflation of L1 estimates into L2 parameters due to group-mean centering into account. In this study, the standardized association of Aggregated TSE with aggregated Reading Comprehension ($\beta = -0.08$, $SE = 0.15$) and Math Achievement ($\beta = 0.004$, $SE = 0.17$) were both non-significant, after controlling for associations of Student-Specific TSE with individual students' Reading Comprehension and Mathematics Achievement at L1. This indicates that the contribution of teachers' feelings of self-efficacy toward an individual student to this student's achievement is not moderated by teachers' average self-efficacy toward the classroom.

Taken together, these results of the model for student-specific measures of TSE suggest that the association between TSE and students' achievement depends on the level of analysis. At the student level, the predictors explained 13.9% of the variance in students' Reading Comprehension and 26.6% of the variance in Mathematics Achievement. At the teacher level, Aggregated TSE explained 7.3% of the variance in aggregated Reading Comprehension and 21.4% in their Math Achievement.

[2] Comparison of this sample with the sample containing whole classrooms revealed similar correlations of General TSE with Reading ($r_{total} = -0.24$; $r_{selected} = -0.24$, *ns*) and Math Achievement ($r_{total} = -0.17$; $r_{selected} = -0.18$, *ns*). This supports once more that the eight randomly selected students from each participating classroom provide a good indication of the classroom mean, see Footnote 1.

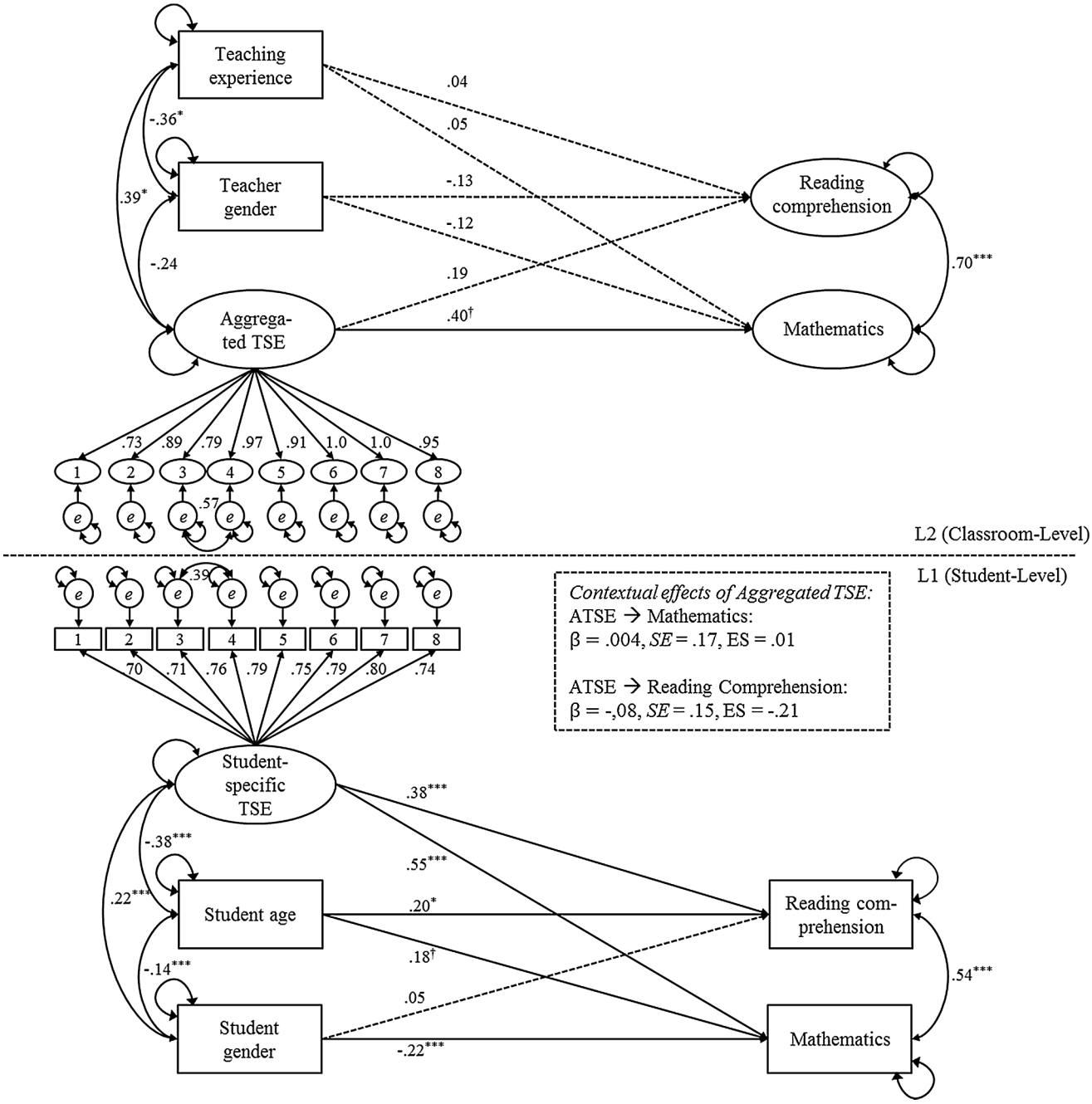

Fig. 2. Doubly latent multilevel structural equation model of aggregated and student-specific TSE. *Note.* Standardized robust maximum likelihood parameter estimates are reported. Dashed lines represent non-significant paths. Gender: 0 = boy/male, 1 = girl/female. TSE = Teacher Self-Efficacy, ATSE = Aggregated Teacher Self-Efficacy, SE = Standard Error, ES = Effect Size. $^{\dagger}p < .10$, $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

4. Discussion

Teachers' self-efficacy beliefs have long been presumed to be associated with students' academic achievement. Yet, empirical evidence suggests that links between TSE and various achievement-related student outcomes do not seem to be as strong and consistent as most researchers have thus far assumed (Klassen et al., 2011; Zee & Koomen, 2016). This study set out to evaluate how different levels of conceptualization and measurement of TSE may affect the association between TSE and students' academic achievement. Guided by the theoretical contentions of Marsh (1986) and Bandura (1997), our results may contribute to the field's understanding of how the conceptualization and measurement of both TSE and achievement at various levels can help to better explain variation in academic outcomes.

4.1. Associations between teachers' self-efficacy and students' achievement

Comparisons of the regression parameters for student-specific, aggregated, and general TSE seemed to suggest that the association between TSE and students' achievement depends on how the multilevel nature of this construct is addressed. This study is probably the first to reveal that teachers' self-efficacy toward individual students at the student level is positively associated with these students' reading comprehension and math performance in class. Following Marsh' (1986) I/E model and Bandura's (1997) social-cognitive notions, such

moderate, positive associations were to be expected, as teachers' student-specific self-efficacy beliefs are tailored to individual students' outcomes and both reflect the same unit of analysis (and theory): the student level (cf. Bandura, 2006; Marsh et al., 2012; Morin et al., 2014).

Interestingly, teachers' student-specific self-efficacy beliefs seemed to be more important to students' math achievement than to their reading comprehension. Specifically, these beliefs accounted for almost twice as much of the variance in students' math achievement than in reading comprehension. This finding resonates well with prior research (e.g., Allinder, 1995; Hines, 2008; Midgley, Feldlaufer, & Eccles, 1989; Throndsen & Turmo, 2013), in which modest positive associations between general TSE and math achievement were found. Possibly, students' math performance may be more affected by instructional and motivational processes than literacy-related tasks, including reading comprehension. The latter tasks are usually more dependent on out-of-school activities, such as print exposure (Mol & Bus, 2011). Especially in the context of middle childhood, it can be assumed that students become more reliant on their own reading strategies and are less dependent on their teachers' instruction. It is possible that highly self-efficacious teachers may display a greater zeal and persistence in teaching mathematics to individual children, as this subject is traditionally perceived by students as more complex and abstract than literacy. Yet, future research is evidently needed to further explore this assumption.

The relationship of the aggregated L2 self-efficacy measure with aggregated levels of mathematics achievement, but not reading comprehension, also appeared to be positive and significant. As we conceived teachers' sense of self-efficacy as a contextual L2 construct, we also tested the contextual effects, specified as the difference between the corresponding L2 and L1 path estimates. The associations at the student and classroom level appeared to be similar in magnitude and the difference among them did not reach statistical significance. This seems to indicate that the contribution of teachers' feelings of self-efficacy toward an individual student to this student's achievement is unlikely to be moderated by the mean level of TSE in the classroom. Hence, at least in our study, we did not find evidence for differences among classrooms in the relationship between student-specific TSE and reading and math achievement.

Overall, the absence of a contextual effect was anticipated, but we are not aware of any theories that may predict it. Yet one possible explanation for the absence of a contextual effect may lie in the extent to which teachers' internal frames of reference are actually affected by their average (aggregated) self-efficacy beliefs. For instance, if teachers feel like their self-efficacy toward one particular student in class is higher than toward another, this may result in them promoting the achievement of this student more than the other. At the same time, however, such positive effects may disappear completely at the classroom level if teachers' student-specific self-efficacy beliefs, which are probably based on relative comparisons of their students, are higher for half of the students in class and lower for the other half. In this case, such average levels of self-efficacy probably do not play a role in aggregated achievement scores (cf. Klein & Kozlowski, 2000; Marsh, 1986).

To some extent, this idea seems to be consistent with the internal/external frame of reference predictions of Marsh and Hau (2004). These predictions suggest that teachers' positive sense of self-efficacy toward one particular student may automatically lead them to judge their self-efficacy toward others in class to be less favorable, thereby potentially affecting individual students' achievement. This may be the case as teachers' internal frame of reference is probably based on their self-efficacy for one particular student as a reference point for judging their self-efficacy toward another (Marsh & Hau, 2004; Parker, Marsh, Lüdtke, & Trautwein, 2013). As such, student-specific beliefs may be more relevant to students' academic achievement than classroom-average TSE, which reflects the average of all the student-level self-efficacy beliefs teachers may hold.

Another possibility is that differences in overall classroom composition, as well as teachers' own characteristics, might affect aggregate TSE and thereby differences across classes. There is some research to suggest that teachers' self-efficacy beliefs may depend on such student features as age, gender, behavior, and level of motivation, and personal teacher characteristics, including gender and teaching experience (Klassen et al., 2009; Raudenbush et al., 1992; Spilt & Koomen, 2009; Zee et al., 2016). If, for example, all classrooms have students with largely similar characteristics, any variation in teachers' student-specific self-efficacy beliefs will probably be canceled out, leading to only small differences in aggregated TSE across classrooms. This may be especially true when teachers have less teaching experience, as these teachers' abilities to recognize the (subtle) needs and behaviors of their students have yet to be developed (Kokkinos & Kargiotidis, 2014; Zee et al., 2016). As such, it is probably hard to find any meaningful relationship with achievement at the classroom level. In contrast, variation in feelings of self-efficacy at the student level, based on internal comparisons, will remain to exist. This might have been the case in our sample, which was relatively homogeneous in terms of students' gender, ethnicity, and socioeconomic status, and teachers' gender. Thus, at least in our study, it seems unlikely that the contribution of teachers' feelings of self-efficacy toward an individual student to this student's achievement depends on the mean level of TSE in the classroom.

Markedly, when using the commonly employed TSES (Tschannen-Moran & Woolfolk Hoy, 2001), the correlation of teachers' general self-efficacy and students' reading and math performance at the between-teacher level appeared to be moderate in magnitude and negative in sign. Thus, although general and aggregated TSE were moderately related to each other, they also clearly differed in their relationship with student achievement at the classroom level. This difference might be somewhat exaggerated due to the use of latent variables that correct for measurement error. This seems particularly to be the case for general TSE, which had a lower reliability than aggregated TSE. Largely consistent with findings from Klassen et al. (2009), the individual factor loading magnitudes of the general TSE construct varied considerably, ranging from 0.34 to 0.86. Such factor loadings may operate to increase the estimate between predictor and criterion constructs, in this study the negative relationship between general TSE and school achievement (e.g., Fornell & Larcker, 1981; Mackenzie, 2001).

Unfortunately, there are no straightforward explanations for the difference in findings between general and aggregated TSE. One possibility is that aggregated TSE might be better tied to teachers' feelings of self-efficacy toward the students in a particular classroom. In contrast, general TSE is probably more affected by some students in the classroom than by others and, in line with the idea of external norm criteria (Marsh, 1986), possibly also influenced by teachers' experiences with other classes. In the latter case, for example, highly self-efficacious teachers' reports of their capability beliefs toward a difficult class may be equal to the reported self-efficacy beliefs of poorly self-efficacious teachers who teach a relatively easy class. Accordingly, the negative correlation between general TSE and academic achievement might have been influenced by the particular combination of classrooms and (characteristics of) teachers in our sample (cf. Marsh, 1986).

The inconsistencies in the difference between general versus student-specific TSE in students' achievement might also be traced back to a relative lack of variation in teachers' general self-efficacy beliefs. Our results as well as evidence from prior research employing the original TSES (e.g., Heneman et al., 2006; Klassen et al., 2009; Tschannen-Moran & Woolfolk Hoy, 2001) has suggested that teachers are likely to respond above the midpoint of the scale, resulting in a considerable degree of range restriction and a lack of variation across teachers. Potentially, such a lack of between-teacher variation in TSE may clarify why the regression of the two achievement measures on teachers' general self-efficacy beliefs is inconsistent with the individual-level coefficient and has a negative sign (cf. Piantadosi et al., 1988). Some empirical studies focusing on general TSE (e.g., Heneman et al., 2006; Reyes et al., 2012) indeed have found that teachers' general self-

efficacy beliefs may negatively contribute to overall levels of student achievement in the classroom.

A last explanation is that small shifts in language between the items in the general and student-specific TSES might have resulted in different results. Although the items of both scales are largely similar in terms of language and meaning, there seems to be a slight lack of parallelism between SE-item 8 of the general TSES ("How much can you do to foster student creativity?") and student-specific TSES ("To what extent can you help this student to explore new things?"). To some extent, these small dissimilarities may explain the difference in findings between general and aggregated TSE.

4.2. Limitations and future directions

The present study's results should be interpreted in the context of several limitations. A first qualification is that we used cross-sectional and correlational data in this study. Although our aim was mainly to illustrate how different levels of conceptualization and measurement of TSE may affect the association between TSE and students' academic achievement, it should be noted that we cannot draw firm conclusions about causal relationships in this study. Of note, teachers' self-efficacy beliefs are generally presumed to be raised and strengthened by experiences of mastery in the classroom, including high-quality student–teacher interactions, students' engagement for their schoolwork, and their academic performance (Bandura, 1997). From this point of view, there is a good possibility that reciprocal relationships existed between teachers' student-specific self-efficacy and students' reading and math achievement (Zee & Koomen, 2016). In any attempt to test this theoretical notion, future researchers are therefore advised to employ longitudinal, cross-lagged (multilevel) designs to disentangle the direction of effects. This may be particularly relevant in a practical sense, since teachers are already getting pressure to tailor their instruction to the specific needs of individual students in the classroom. By employing longitudinal doubly latent ML-SEM, researchers might be able to clarify how student-specific and general TSE may be at play in influencing teachers' instructional choices and behaviors in class.

A second qualification is that we used instruments to capture various domains of teachers' self-efficacy. Although these domains are presumed to reflect the breadth of elementary teachers' daily activities, there is some evidence that teachers may also feel more or less self-efficacious depending upon the specific subject area they have to teach, including literacy and science (e.g., Riggs & Enochs, 1990; Tschannen-Moran & Johnson, 2011). To further elucidate the link between TSE and students' achievement in reading and mathematics, it may be helpful to take account of teachers' subject-specific self-efficacy as well.

Additionally, the results of this study may not be generalizable to the total population of students and teachers in the elementary grades. To be more precise, our sample included a large amount of female teachers who appeared to be relatively experienced in teaching. Moreover, the students in this sample came from relatively privileged backgrounds and were predominantly Dutch. Accordingly, our sample was relatively homogenous in terms of students' and teachers' backgrounds. To increase the generalizability of the current study's results, inclusion of teachers and students from more heterogeneous milieus may probably warrant consideration in future studies.

Last, it should be noted that our sample size, and the number of clusters in particular, was rather small. Simulation results of Lüdtke et al. (2009) have shown that doubly latent ML-SEM generally requires at least 50 L2 units with at least 10–15 participants per unit and that smaller samples may lead to model non-convergence and estimation errors. Our study consisted of 49 teachers, reporting on eight students on average. Yet, we should also note that ML-SEM was used under very favorable conditions (Lüdtke et al., 2009). Factor loadings of the indicators of our latent variables were high. Additionally, there was a good level of agreement between the student-specific TSE ratings and achievement scores provided by the eight students in each participating classroom. This has not only resulted in excellent reliabilities of aggregated scores, but also indicates that randomly selecting eight students from each teachers' classroom is sufficient to obtain consistent results. Also, these relatively high reliability indexes have probably accounted for the small sample size, and overcome estimation problems in our study. Nevertheless, larger samples may warrant consideration in future studies on the multilevel nature of TSE.

5. Conclusion

The findings of this study seem to underscore the importance of addressing the inherently multilevel nature of the concept of teachers' self-efficacy and recognizing that predictors and outcomes at the between-teacher level frequently measure different constructs than their namesakes at the student level. Whereas student-specific TSE was positively associated with students' academic achievement at L1, general TSE appeared to be negatively related to reading and math achievement at L2. Moreover, our results indicated that the contribution of teachers' feelings of self-efficacy toward an individual student to this student's achievement does not seem to be dependent on teachers' average, or aggregated self-efficacy in the classroom. Taken together, this knowledge may be a first step forward in spurring further understanding of the complex, potentially reciprocal relationship between teachers' sense of self-efficacy and students' academic achievement in elementary school, as well as the underlying mechanisms that explain this association.

Appendix A

See Table A1.

Table A1
General and student-specific TSES items.

Domain	Item	Teachers' general self-efficacy	Teachers' student-specific self-efficacy
IS	1	To what extent can you provide an alternative explanation or example when students are confused?	To what extent can you provide an alternative explanation or example when this student is confused?
IS	2	To what extent can you gauge student comprehension of what you have taught?	To what extent can you gauge this student's comprehension of what you have taught?
IS	3	How much can you do to adjust your lessons to the proper level for individual students?	How well can you adjust your lessons to the proper level for this student?
IS	4	How well can you provide appropriate challenges for very capable students?	How well can you provide appropriate challenges for this student?
SE	5	How much can you do to get students to believe they can do well in schoolwork?	How well can you get this student to believe he/she can do well in schoolwork?
SE	6	How much can you do to help your students value learning?	To what extent can you help this student to value learning?
SE	7	How much can you do to help your students think critically?	How well can you help this student to think critically?
SE	8	How much can you do to foster student creativity?	To what extent can you help this student to explore new things?

Note. IS = Instructional strategies; SE = Student engagement.

References

Allinder, R. M. (1995). An examination of the relationship between teacher efficacy and curriculum-based measurement and student achievement. *Remedial & Special Education, 16*, 247–255.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84*, 191–215. https://doi.org/10.1037/0033-295X.84.2.191.

Bandura, A. (1997). *Self-efficacy: The exercise of control.* New York, NY: W.H. Freeman.

Bandura, A. (2006). Guide for constructing self-efficacy scales. In F. Pajares, & T. Urdan (Vol. Eds.), *Adolescence and education: Self-efficacy and adolescence: Vol. 5*, (pp. 307–337). Greenwich, CT: Information Age.

Bandura, A. (2012). On the functional properties of perceived self-efficacy revisited. *Journal of Management, 38*, 9–44.

Bentler, P. M. (1992). On the fit of models to covariances and methodology to the Bulletin. *Psychological Bulletin, 112*, 400–404. https://doi.org/10.1037/0033-2909.112.3.400.

Bing, M. N., Whanger, J. C., Davison, H. K., & VanHook, J. B. (2004). Incremental validity of the frame-of-reference effect in personality scale scores: A replication and extension. *Journal of Applied Psychology, 89*, 150–157. https://doi.org/10.1037/0021-9010.89.1.150.

Brown, R. L. (1989). Using covariance modeling for estimating reliability on scales with ordered polytomous variables. *Educational and Psychological Measurement, 49*, 385–398.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen, & J. S. Long (Eds.). *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Cantrell, S. C., Almasi, J. F., Carter, J. C., & Rintamaa, M. (2013). Reading intervention in middle and high schools: Implementation fidelity, teacher efficacy, and student achievement. *Reading Psychology, 34*, 26–58. https://doi.org/10.1080/02702711.2011.577695.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*, 284–290. https://doi.org/10.1037/1040-3590.6.4.284.

Eberle II, W. M. (2011). Teacher self-efficacy and student achievement as measured by North Carolina reading and math end-of-grade tests. Retrieved from ProQuest Dissertations and Theses database (UMI No. 3462062).

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling, 8*, 430–457. https://doi.org/10.1207/S15328007SEM0803_5.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*, 39–50.

Geerlings, J., Thijs, J., & Verkuyten, M. (2017). Teaching in ethnically diverse classrooms: Examining individual differences in teacher self-efficacy. *Journal of School Psychology, 67*, 134–147. https://doi.org/10.1016/j.jsp.2017.12.001.

Guo, Y., McDonald Connor, C., Yang, Y., Roehring, A. D., & Morrison, F. J. (2012). The effects of teacher qualification, teacher self-efficacy, and classroom practices on fifth graders' literacy outcomes. *The Elementary School Journal, 113*, 3–24. https://doi.org/10.1086/665816.

Hardré, P. L., & Sullivan, D. W. (2008). Teacher perceptions and individual differences: How they influence rural teachers' motivating strategies. *Teaching and Teacher Education, 24*, 2059–2075. https://doi.org/10.1016/j.tate.2008.04. 007.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis, 29*, 60–87. https://doi.org/10.3102/0162373707299706.

Heneman, H. G., Kimball, S., & Milanowski, A. (2006). *The Teacher Sense of Efficacy Scale: Validation evidence and behavioral prediction (WCER Working Paper No. 2006-7).* Madison, WI: Wisconsin Center for Education Research.

Hines, M. T. (2008). The interactive effects of race and teacher self-efficacy on the achievement gap in school. *National Forum of Multicultural Issues Journal, 7*, 1–11.

Janssen, J., Verhelst, N., Engelen, R., & Scheltens, F. (2010). *Wetenschappelijke verantwoording van de toetsen LOVS Rekenen-Wiskunde voor groep 3 tot en met 8 [Scientific justification of the LOVS Mathematics tests for grades 1 to 6].* Arnhem, the Netherlands: CITO.

Klassen, R. M., Bong, M., Usher, E. L., Chong, W. H., Huan, V. S., Wong, I. Y., & Georgiou, T. (2009). Exploring the validity of a teachers' self-efficacy scale in five countries. *Contemporary Educational Psychology, 34*, 67–76. https://doi.org/10.1016/j.cedpsych.2008.08.001.

Klassen, R. M., & Tze, V. M. (2014). Teachers' self-efficacy, personality, and teaching effectiveness: A meta-analysis. *Educational Research Review, 12*, 59–76. https://doi.org/10.1016/j.edurev.2014.06.001.

Klassen, R. M., Tze, V. M., Betts, S. M., & Gordon, K. A. (2011). Teacher efficacy research 1998–2009: Signs of progress or unfulfilled promise? *Educational Psychological Review, 23*, 21–43. https://doi.org/10.1007/s10648-010-91418.

Klein, K. J., & Kozlowski, S. W. (2000). From micro to meso: Critical steps in conceptualizing and conducting multilevel research. *Organizational Research Methods, 3*, 211–236.

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford.

Kokkinos, C. M., & Kargiotidis, A. (2014). Rating students' problem behaviour: The role of teachers' individual characteristics. *Educational Psychology, 36*, 1516–1532. https://doi.org/10.1080/01443410.2014.993929.

Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2 × 2 taxonomy of multilevel latent contextual models: Accuracy–bias trade-offs in full and partial error correction models. *Psychological Methods, 16*, 444–467. https://doi.org/10.1037/a0024376.

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

Lüdtke, O., Robitzsch, A., Trautwein, U., & Kunter, M. (2009). Assessing the impact of learning environments: How to use student ratings of classroom or school characteristics in multilevel modeling. *Contemporary Educational Psychology, 34*, 120–131.

Lumpe, A., Czerniak, C., Haney, J., & Beltyukova, S. (2012). Beliefs about teaching science: The relationship between elementary teachers' participation in professional development and student achievement. *International Journal of Science Education, 34*, 153–166. https://doi.org/10.1080/09500693.2010.551222.

Maassen, G. H., & Bakker, A. B. (2001). Suppressor variables in path models: Definitions and interpretations. *Sociological Methods & Research, 30*, 241–270. https://doi.org/10.1177/0049124101030002004.

MacKenzie, S. B. (2001). Opportunities for improving consumer research through latent variable structural equation modeling. *Journal of Consumer Research, 28*, 159–166.

Marsh, H. W. (1986). Verbal and mathematics self-concepts: An internal/external frame of reference model. *American Educational Research Journal, 23*, 129–149.

Marsh, H. W. (1991). The failure of high ability high schools to deliver academic benefits: The importance of ASC and educational aspirations. *American Educational Research Journal, 28*, 445–480.

Marsh, H. W., & Hau, K. (2004). Explaining paradoxical relations between academic self-concepts and achievements: Cross-cultural generalizability of the internal/external frame of reference predictions across 26 countries. *Journal of Educational Psychology, 96*, 56–67.

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J. S., Abduljabbar, A. S., & Köller, O. (2012). Classroom climate and contextual effects: Conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist, 47*, 106–124. https://doi.org/10.1080/00461520.2012.670488.

Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B. O., & 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.

Martin, N. K., Sass, D. A., & Schmitt, T. A. (2012). Teacher efficacy in student engagement, instructional management, student stressors, and burnout: A theoretical model using in-class variables to predict teachers' intent-to-leave. *Teaching and Teacher Education, 28*, 546–559. https://doi.org/10.1016/j.tate.2011.12.003.

Midgley, C., Feldlaufer, H., & Eccles, J. (1989). Change in teacher efficacy and student self- and task-related beliefs in mathematics during the transition to junior high school. *Journal of Educational Psychology, 81*, 247–258. https://doi.org/10.1037/0022-0663.81.2.247.

Mojavezi, A., & Poodineh Tamiz, M. (2012). The impact of teacher self-efficacy on the students' motivation and achievement. *Theory and Practice in Language Studies, 2*, 483–491. https://doi.org/10.4304/tpls.2.3.483-491.

Mol, S. E., & Bus, A. G. (2011). To read or not to read: A meta-analysis of print exposure from infancy to early adulthood. *Psychological Bulletin, 137*, 267–296. https://doi.org/10.1037/a0021890.

Morin, A. J., Marsh, H. W., Nagengast, B., & Scalas, L. F. (2014). Doubly latent multilevel analyses of classroom climate: An illustration. *The Journal of Experimental Education, 82*, 143–167. https://doi.org/10.1080/00220973.2013.769412.

Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Nie, Y., Tan, G. H., Liau, A. K., Lau, S., & Chua, B. L. (2013). The roles of teacher efficacy in instructional innovation: Its predictive relations to constructivist and didactic instruction. *Educational Research for Policy and Practice, 12*, 67–77. https://doi.org/10.1007/s10671-012-9128-y.

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

Piantadosi, S., Byar, D. P., & Green, S. B. (1988). The ecological fallacy. *American Journal of Epidemiology, 127*, 893–904.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods, Vol. 1.* London: Sage.

Raudenbush, S. W., Rowan, B., & Cheong, Y. F. (1992). Contextual effects on the self-perceived efficacy of high school teachers. *Sociology of Education, 65*, 150–167.

Reyes, M. R., Brackett, M. A., Rivers, S. E., White, M., & Salovey, P. (2012). Classroom emotional climate, student engagement, and academic achievement. *Journal of Educational Psychology, 104*, 700–712. https://doi.org/10.1037/a0027268.

Riggs, I. M., & Enochs, L. G. (1990). Toward the development of an elementary teacher's science teaching efficacy belief instrument. *Science Education, 74*, 625–637. https://doi.org/10.1002/sce.3730740605.

Ross, J. A. (1992). Teacher efficacy and the effects of coaching on student achievement. *Canadian Journal of Education, 17*, 51–65. https://doi.org/10.2307/1495395.

Sanders, W. L., Wright, S. P., & Horn, S. P. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education, 11*, 57–67. https://doi.org/10.1023/A:1007999204543.

Shin, T., Davidson, M. L., & Long, J. D. (2009). Effects of missing data methods in structural equations modeling with nonnormal data. *Structural Equation Modeling, 16*, 70–98. https://doi.org/10.1080/10705510802569918.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* London, UK: Sage Publishers.

Spilt, J. L., & Koomen, H. M. Y. (2009). Widening the view on teacher–child relationships: Teachers' narratives concerning disruptive versus nondisruptive children. *School Psychology Review, 38*, 86–101.

Thoonen, E. E. J., Sleegers, P. J. C., Oort, F. J., Peetsma, T. T. D., & Geijsel, F. P. (2011). How to improve teaching practices: The role of teacher motivation, organizational factors, and leadership practices. *Educational Administration Quarterly, 47*, 496–536. https://doi.org/10.1177/0013161X11400185.

Throndsen, I., & Turmo, A. (2013). Primary mathematics teachers' goal orientations and student achievement. *Instructional Science, 41*, 307–322. https://doi.org/10.1007/s11251-012-9229-2.

Tournaki, N., & Podell, D. M. (2005). The impact of student characteristics and teacher efficacy on teachers' predictions of student success. *Teaching and Teacher Education, 21*, 299–314. https://doi.org/10.1016/j.tate.2005.01.003.

Tschannen-Moran, M., & Johnson, D. (2011). Exploring literacy teachers' self-efficacy beliefs: Potential sources at play. *Teaching and Teacher Education, 27*, 751–761. https://doi.org/10.1016/j.tate.2010.12.005.

Tschannen-Moran, M., & Woolfolk Hoy, A. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education, 17*, 783–805. https://doi.org/10.1016/S0742-051X(01)00036-1.

Tschannen-Moran, M., Woolfolk Hoy, A., & Hoy, W. K. (1998). Teacher efficacy: Its meaning and measure. *Review of Educational Research, 68*, 202–248. https://doi.org/10.3102/00346543068002202.

Tsouloupas, C. N., Carson, R. L., Matthews, R., Grawitch, M. J., & Barber, L. K. (2010). Exploring the association between teachers' perceived student misbehaviour and emotional exhaustion: The importance of teacher efficacy beliefs and emotion regulation. *Educational Psychology, 30*, 173–189. https://doi.org/10.1080/014434109 03494460.

Weekers, A., Groenen, I., Kleintjes, F., & Feenstra, H. (2011). *Wetenschappelijke verantwoording papieren toetsen Begrijpend lezen voor groep 7 en 8 [Scientific justification of Reading Comprehension tests for grades 5 and 6]*. Arnhem, the Netherlands: CITO.

Wheatley, K. F. (2005). The case for reconceptualizing teacher efficacy research. *Teaching and Teacher Education, 21*, 747–766. https://doi.org/10.1016/j.tate.2005.05.009.

Zee, M., de Jong, P. F., & Koomen, H. M. Y. (2016). Teachers' self-efficacy in relation to individual students with a variety of social-emotional behaviors: A multilevel investigation. *Journal of Educational Psychology, 108*, 1013–1027. https://doi.org/10.1037/edu0000106.

Zee, M., de Jong, P. F., & Koomen, H. M. Y. (2018). Omgaan met verschillende soorten gedrag in de klas: De rol van leerkracht self-efficacy. [Dealing with diversity in the classroom: The role of teacher self-efficacy]. *Kind en Adolescent, 1*, 1–21. https://doi.org/10.1007/s12453-017-0162-7.

Zee, M., & Koomen, H. M. Y. (2016). Teacher self-efficacy and its effects on classroom processes, student academic adjustment, and teacher well-being A synthesis of 40 years of research. *Review of Educational Research, 86*, 981–1015. https://doi.org/10.3102/0034654315626801.

Zee, M., Koomen, H. M. Y., Jellesma, F. C., Geerlings, J., & de Jong, P. F. (2016). Inter-and intra-individual differences in teachers' self-efficacy: A multilevel factor exploration. *Journal of School Psychology, 55*, 39–56. https://doi.org/10.1016/j.jsp.2015.12.003.