RUNNING HEAD: Multilevel Analysis of Climate Effects

Doubly Latent Multilevel Analyses of Classroom Climate: An illustration

Alexandre J. S. Morin*[1], Herbert W. Marsh*[1,2,3], Benjamin Nagengast[4], L. Francesca Scalas[5]

[1] Centre for Positive Psychology and Education, University of Western Sydney, Australia; [2] University of Oxford, UK;[3] King Saud University, Saudi Arabia; [4] Center for Educational Science and Psychology (CESP), Department of Education, University of Tübingen, Germany; [5] Università degli Studi di Cagliari, Italy

* The first two authors (A. J. S. M. and H. W. M.) contributed equally to this article and their order was determined at random: both should thus be considered first authors.

This is the final prepublication version of:

Morin, A.J.S., Marsh, H.W., Nagengast, B., & Scalas, L.F. (2014). Doubly Latent Multilevel Analyses of Classroom Climate: An illustration. *Journal of Experimental Education, 82* (2), 143-167

Acknowledgements

The authors are very grateful to the authors of the original Fast, Lewis, Bryant, Bocian, Cardullo, Rettig and Hammond (2010) study, and especially to Lisa A. Fast and James L. Lewis, who were particularly helpful at the initial stage of this research. We appreciate d their willingness to provide us with the original and expanded data set and to help us set up the analyses. We note, however, that the interpretations of the data are the responsibility of the authors of this study and may not represent those of Fast et al. The revision of this manuscript was conducted while the first author was a visiting scholar at the University of Cagliari.

Corresponding author
Corresponding author:
Alexandre J.S. Morin
Centre for Positive Psychology and Education
University of Western Sydney
Locked Bag 1797, Penrith, NSW 2751, Australia
E-mail: A.Morin@uws.edu.au

Abstract

Many classroom climate studies suffer from two critical problems. They: (a) treat climate as a student-level (L1) variable in single-level analyses instead of a classroom-level (L2) construct in multilevel analyses; and (b) rely on manifest-variable models rather than latent-variable models that control measurement error at L1 and L2, and sampling error in the aggregation of L1 ratings to form L2 constructs. Based on an analysis of 2541 students in Grades 5 or 6 from 89 classrooms, we demonstrate doubly-latent multilevel structural equation models that overcome both these problems. The results show that L2 classroom climate (a higher-order factor representing classroom mastery goal orientation, challenge, and teacher caring) had positive effects on self-efficacy and achievement. We conclude with a discussion of related issues (e.g. the meaning of L2 constructs versus L1 residuals, the dimensionality of climate constructs at L2) and guidelines for future research.

Keywords: Classroom climate, self-efficacy, achievement, doubly-latent multilevel structural equation models.

Doubly Latent Multilevel Analyses of Classroom Climate: An illustration

Interesting substantive problems lead to cutting-edge methodological developments, which, in principal should lead to greater insights into substantive problems. (Kaplan, 2009, p.9)

Recent years have seen a dramatic rise in the development of advanced methodology for educational research. This creates new research opportunities as innovations enable researchers to penetrate previously inaccessible research problems, revisit classic unresolved issues with stronger tools, and address new research questions. Addressing this issue, Marsh and Hau (2007; also see Borsboom, 2006) argued for that there was a need for substantive-methodological synergies in educational research. More precisely, they note that some of the best methodological research is based on the development of creative methodological solutions to problems that stem from substantively complex research questions. In turn, new methodologies provide new ways of addressing complex substantive issues. However, these substantive-methodological synergies require substantive researchers to keep pace with new developments and to participate in on-going high-level statistical training and self-development. The converse is also true where quantitative researchers must continue to engage with problems that have practical, substantive, and theoretical significance. The present investigation is one such substantive-methodological synergy (Marsh & Hau, 2007), bringing to bear new, strong, and evolving methodology in order to evaluate classroom climate effects and is specifically designed to illustrate these methods to applied researchers not familiar with them.

In educational research, climate studies evaluate whether school, classroom, or teacher (group-level, L2) characteristics contribute to the prediction of students' (individual-level, L1) outcomes (e.g., achievement, self-concept, engagement, persistence) beyond what can be explained by other individual characteristics of students. In many studies, L2 constructs have been based on the aggregation of L1 student-level variables (e.g., ratings of classroom organization or evaluations of teacher enthusiasm used to form classroom-level climate variables). This general strategy has been at the heart of educational research (e.g., school/teacher effectiveness studies; value-added models; classroom/school climate) and generalizes to other psychological and social science research. However, reviews of this research (e.g., Lau & Nie, 2008; Marsh, et al., 2012; Miller, 2006; Miller &

Murdock, 2007) identify at least two critical problems that are common to many classroom climate studies and that are the focus of the present investigation: the failure to consider the appropriate level of analysis and the absence of proper control for measurement and sampling error.

Level of Analysis and Types of L2 Constructs

The most serious problem is the failure to treat *classroom* climate as a classroom-level construct rather than an individual student-level construct. Methodologically, this issue of the relevant unit of analysis was clearly articulated more than a quarter of a century ago in Cronbach's (1976, p. 18) seminal article on multilevel issues in classroom climate research, where he argued that:

> *The purpose of the LEI [Learning Environment Inventory] is to identify differences among classrooms. For it, then, studies of scale homogeneity or scale intercorrelation should be carried out with the classroom group as unit of analysis. Studying individuals as perceivers within the classrooms could be interesting, but is a problem quite separate from the measurement of environments.*

Despite well-articulated theoretical and statistical rationales following from Cronbach (e.g., Lau & Nie, 2008; Lüdtke, Robitzsch, Trautwein & Kunter, 2009; Marsh, et al., 2012; Miller, 2006; Miller & Murdock, 2007; Papaioannou et al., 2004), there still exists widespread confusion in educational research about the appropriate nature of data, design, statistical models, and interpretations of classroom climate research. For example, in their review of classroom goal structures, Miller (2006; Miller & Murdock, 2007) found that 16 of 31 studies did not consider any classroom level of analysis.

If researchers are interested in the effects of L2 (classroom, teacher, or school) variables, then the appropriate unit of analysis should be an L2 unit. [For present purposes we focus on classroom effects, but the approach we demonstrate is also relevant to studies where school or teacher is the L2 unit]. Hence, evaluation of the effects of classroom climate should be based on L2 classroom-level constructs formed by the aggregations of ratings by students, not on L1 student-level responses. In fact, in evaluating classroom climate, this interest for what is occurring at the classroom level is made obvious from the fact that students are usually asked to directly rate characteristics of the L2 classroom that are common to all students and not to rate some specific personal characteristics of themselves. That is, the referent, i.e. the external "objective" reality being assessed, is the same for all

students in the class. In fact, most of the commonly used measures of classroom climate are implicitly based on an interest for the L2 level that should be made explicit in the statistical model that is used.

In fact, interpreting student-level (L1) perceptions of classroom climate as if they reflect classroom-level (L2) climate is a classic example of the ecological fallacy identified by Robinson (1950) more than 60 years ago, well before Cronbach (1976). This classical and very common mistake in educational research involves the implicit assumption that the effects observed at one level generalize to another (also see, Marsh et al., 2009; Schwartz, 1994). Thus, even if one was to argue that climate ratings taken at the individual level had meaning in themselves as reflecting some relevant characteristics of the individual (i.e., like gender or academic achievement), analysing them at a single level (i.e., student-level) confounds the effects of the individual student and the classroom (or school) and implicitly assume that both effects are the same.

A vivid illustration of the ecological fallacy comes from classic research on the big-fish-little-pond effect (Marsh, 2007a). This research shows that achievement at the individual student level has a positive effect on academic self-concept (the brighter I am the better my academic self-concept), but school- or classroom-average achievement has a negative effect on academic self-concept (the brighter my classmates, the lower my academic self-concept). In this case, a single-level analysis at the student level leads to underestimation of the relation between achievement and self-concept at the individual student level and completely ignores the big-fish-little-pond effect at the school level, due to the conflation of strong student-level positive effects with smaller negative school-level effects. For climate research were students are asked to directly rate the L2 reality they are exposed to, we can logically expect these problems to be even more serious since failure to analyse the data at the proper level of analysis would lead the researcher to conclude that the effects are located at the individual level when in fact they are located at the classroom level.

For example, Papaioannou, Marsh, and Theodorakis (2004) simultaneously evaluated the L2 classroom competitiveness climate in physical education classes (i.e., students were asked to rate the class) and the L1 personal competitive orientation of the students (i.e., students were asked to rate themselves). They used these ratings to test the cross-level interaction effects that competitive students would be advantaged in more competitive classes, but found no support for this “matching”

hypothesis. Rather, their results showed that the L1 and L2 constructs tapped into different, and independent, realities. More recently, Marsh et al. (2012) showed that once classroom climate was appropriately modeled as an L2 construct, the L1 residual ratings of classroom climate (corresponding to deviations between the way individual students rated their class and the average classroom rating) had no remaining effects on student's levels of mathematic self-concept and achievement.

Although we emphasize *climate* constructs in the present study, two types of L2 constructs can be distinguished based on the aggregation of L1 ratings (e.g., Marsh et al., 2009, 2012; Skrondal & Laake, 2001): contextual and climate constructs. Classroom-level L2 contextual variables are based on aggregates of L1 ratings that are specific to the person being assessed and meaningful in themselves (e.g. gender, or students' levels of math self-efficacy and achievement, such as those used by Fast et al., 2010). In this respect, the students in the same classroom are not 'interchangeable' since the class composition represents a true aggregate of the individual characteristics of the students composing it. Thus, the L1 measures used to construct contextual variables are potentially important in their own right and may have a distinct meaning from the associated L2 contextual variables (e.g. class average levels of achievement). Because L1 ratings are meaningful in their own right, the effects of L2 contextual constructs need to be estimated after controlling for the students' differences in the corresponding L1 variable (e.g., the effects of class-average math self-efficacy after controlling for the effects of individual student math self-efficacy; Marsh et al., 2009, 2012, also see Appendix two for additional technical considerations). In this case, failure to properly disentangle these effects can lead to biases similar to those identified in the Big-Fish-Little-Pond effect where two meaningful, yet different, effects got conflated.

Conversely, classroom-level L2 climate variables are based on aggregates of student ratings where the group is the referent and not the individual. That is, each student directly rates the L2 construct on items having the classroom as the referent and not on characteristics specific to that individual student. Thus, aggregated climate variables assume that scores for all students within the same classroom reflect the same underlying L2 construct and thus, that students within the same classroom are theoretically interchangeable (i.e. all students within a class are rating the same classroom climate) and rate the L2 construct directly. From this statistical perspective, differences

among students within the same classroom represent a source of unreliability in the L2 construct. Thus, for L2 climate variables, the L2 climate effect is the effect of the aggregated L2 construct—not the effect of the L1 ratings used to construct the L2 climate measure. Hence, the properly disaggregated L1 component reflects differences in perceptions of this L2 construct by students within the same class and a source of unreliability of the aggregated class ratings, not individual differences among students in the classroom climate construct being assessed. In this case, failure to properly model climate at L2 may lead to erroneous conclusion as to the location, or source, of the effect: individual perceptions or classroom characteristics, as shown in Marsh et al. (2012) illustration.

When studying classroom climate constructs, the referent of the items should be the L2 classroom, in that each student in the class rates some aspect of the classroom rather than some individual characteristic of the student making the rating. Furthermore, the class referent needs to, as is typically the case, be made explicit. For example, the widely used Pattern of Adaptive Learning Scales (PALS; Midgley, 2002) include dimensions related to classroom mastery goal structure ('In our **class**, how much you improve is really important'), classroom performance-approach goal structure ('In our **class**, getting good grades is the main goal'), and performance-avoidance goal structure ('In our **class**, showing others that you are not bad at class work is really important') where the explicit referent is always the classroom. When the referent is an L2 unit, the L1 ratings assess an L2 classroom construct, not individual student characteristics. In fact, even when one claims to be interested in the study of individual students' perceptions of the classroom climate (e.g., Ciani, Middleton, Summers, & Sheldon, 2010; Fast, et al., 2010), the ratings still reflect—perhaps substantially—the L2 classroom characteristics rather than individual differences in the way specific students view climate. Thus, in single level analyses that do not control for L2 effects, L2 classroom climate and residual L1 effects that reflect how the perceptions of individual students within a class differ from the class average are confounded. As previously shown, this confounding may result in systematic biases (e.g. Marsh et al., 2009 , 2012; Papaioannou et al., 2004 ; for further discussion of these biases see Miller & Murdock, 2007). As emphasized by Marsh et al. (2012) these two components of classroom climate ratings (i.e. the L2 shared agreement among students within the same classroom and the L1 residual variances at the level of the individual students) are automatically disentangled in appropriate multilevel models.

However, the residual L1 climate ratings have no substantive meaning in relation to the interpretation of the L2 climate effects and only need to be considered in the model in order to properly control for unreliability and sampling error in the aggregation of individual ratings into L2 constructs.

From this perspective there should, optimally, be good agreement among students within the same class; a complete lack of agreement would suggest that the climate variable is completely unreliably measured and probably should not be considered further as a measure of classroom climate. The L1 ratings of climate are thus important in terms of estimating agreement among students within the class and forming the L2 aggregates. Indeed, a highly reliable measure of L2 classroom climate requires good agreement among students within each class and relatively little residual variance at the L1 student level, beyond what can be explained by the L2 climate factor. If there is little agreement among students within the same class, then their ratings do not reflect *classroom* climate.

The residual L1-ratings (i.e., individual student ratings after controlling for the corresponding class-average ratings) might or might not be systematically related to other L1 constructs. However, the L1-ratings (residualized or not) do not represent the L2 classroom climate, nor do they represent individual students' characteristics. The residual L1-ratings represent unique perceptions of each student that are not explained by the shared perceptions of different students and reflect a form of inter-individual measurement error in the assessment of L2 constructs. Marsh et al. (2012; also see Papaioannou et al., 2004) have noted that these residual L1 ratings of climate might reflect systematic method effects (e.g., positive or negative response biases), or the presence of subclimates/subcultures within the classroom. In this sense, they may also have a substantive role in the interpretation of results. However, they caution that any such substantive interpretations should be based on a clear theoretical and statistical rationale, which includes disentangling L1 and L2 effects and appropriately interpreting the residualized L1 ratings. One thing is clear however: the L1 residual ratings represent individual differences in the perception of an L2 construct, not individual characteristics per se. If applied researchers want to measure characteristics of individual students, then the referent should be L1 individual students (e.g., My math teacher cares about how I feel) instead of—or in addition to—the L2 classroom (e.g., Our math teacher cares about how we feel).

Measurement and Sampling Error

The second critical problem is that many classroom climate studies use manifest variables based on aggregated scale scores rather than relying on latent variable methodologies such as structural equation modeling (SEM). SEM would allow classroom climate studies to control for measurement error. Although measurement error attenuates the sizes of correlations in predictable ways, effects of measurement error are not so obvious in multilevel path models, where the direction of the bias can be positive or negative in different parts or levels of the model (e.g. Lüdtke, Marsh, Robitzsch, Trautwein, 2011; Lüdtke, Marsh, Robitzsch, Trautwein, Asparouhov, & Muthén, 2008; Marsh et al., 2010). As noted by Marsh, Lüdtke et al. (2009 , 2012; Lüdtke et al., 2008; 2011), the recent development of doubly latent multilevel models (ML-SEM) allow for the use of multiple indicators to control measurement error at both the individual student and classroom levels, while also controlling for sampling error in the aggregation of responses by individual students to represent classroom level constructs. These models are thus doubly latent, in relation to: (a) ***Measurement error***, as in traditional factor analyses, since they incorporate multiple indicators (e.g., items) of the constructs at both the L1 student level and the L2 classroom level; (b) ***Sampling error***, as in traditional multilevel analyses, since they incorporate L1 scores for different students in the same class as multiple indicators of latent class-level constructs that are corrected for sampling error in the aggregation of L1 scores to form latent L2 constructs. Although we target an audience of methodologically informed applied researchers in the present illustration of ML-SEM models, we refer the more technically-oriented readers to previous technical publications for more formal presentations of ML-SEM models and the way specific controls for measurement and sampling error are implemented (e.g., Muthén, & Asparouhov, 2011; Lüdtke et al., 2008; 2011; Marsh et al., 2009).

Why has the inherently multilevel structure of classroom climate been inappropriately represented in so many applied educational psychology publications? We surmise that the answer is a complex interplay between substantive and methodological issues. Historically, educational psychology focused on single-level models of manifest variables—ignoring the inherent multilevel structure of most educational data, sampling error in forming the L2 units when these where even considered, and measurement error at L1 and L2. Based in part on this single-level thinking, L1 student ratings of classroom climate are sometimes conceived of—inappropriately—as L1 student

characteristics rather than as L2 classroom characteristics, or as a confounding of these L1 and L2 influences. Also, complications arise when the main independent variable (classroom climate) is an L2 construct, but the main outcome variables are L1 constructs. However, even in this case, the primary focus should still be on L2 classroom climate (based aggregating students' responses within the same classroom) and its effects on other constructs. Indeed, if the effect of classroom climate purportedly occurs at the classroom level, then this effect should first and foremost be studied at the classroom level (Preacher, Zhang, & Zyphur, 2011; Preacher, Zyphur, & Zhang, 2010). It is only then that it becomes possible to clearly investigate the role of the residualized L1 ratings as possible indicators of individual student differences in perceptions of the L2 classroom—if one can find a reasonable argument to justify focusing on these residuals. Furthermore, appropriate conceptualization of classroom climate as a classroom-level construct requires representative samples from a large number of classrooms, which are beyond the resources of many applied researchers. Methodologically, innovations in confirmatory factor analysis, structural equation models and multilevel modeling are now part of the arsenal of educational psychologists. However, the synergy of these two methodologies—and the appropriate software—has been slow to evolve (Marsh, et al., 2009, 2012).

The Present Investigation

In order to illustrate evolving developments in classroom climate research, we analyzed a data set that was used in a recent classroom climate study published in the *Journal of Educational Psychology* (Fast et al., 2010). Fast et al. (2010) investigated the role of classroom climate on elementary school students' levels of math self-efficacy and achievement. Their conclusions were that 'students who perceived their classroom environments as more caring, challenging, and mastery oriented had significantly higher levels of math self-efficacy and higher levels of math self-efficacy positively predicted math performance' (p. 729), supporting a fully mediated hypothesis. The model tested by Fast et al. (2010) is presented in Figure 1, with controls represented by dashed lines.

Fast et al. (2010) presented an analysis of 1163 students (from 88 classrooms) measured twice: once when they were in fourth or fifth grade and once again one year later when they were in fifth or sixth grade. They conducted analyses based on the second time (Time 2, T2), using data from the previous school year as controls (Time 1, T1). This study has many admirable qualities (e.g., the large

sample of many students from many schools, preliminary CFAs of L1 classroom climate ratings to ascertain the dimensionality of the climate measure, and the application of two-level models to control for the nesting of students into classes) that contributed to its publication in a leading educational psychology journal. Nevertheless, like many classroom climate studies, it suffered from both of the problems outlined above: (a) L1 climate ratings by individual students were the basis of interpretations of classroom climate, even though the L1 residual variances were not disentangled from the L2 classroom aggregations of the L1 ratings, thus resulting in a conflation of L1 inter-individual differences in perception of the classroom climate with aggregated L2 classroom characteristics; (b) sampling error in forming the L2 unit was ignored; and (c) although many of the constructs were based on multiple indicators, and preliminary CFAs were conducted, the classroom climate analysis was based on manifest variables that failed to control for measurement error. For these reasons, it was an ideal study with which to investigate a ML-SEM approach to classroom climate research: we are grateful to the authors of the original study for their generosity, their willingness to provide us with the data and their assistance in early analyses of the present investigation.

Methods

The Fast et al. (2010) study was based on responses by 1163 participants who were selected from a larger pool of students, excluding students who did not have data for both years of the study or who had missing data on any of the control or achievement variables (i.e. a quasi-listwise deletion strategy). However, ML-SEM models require large sample sizes and provide more accurate interpretations when each classroom is based on a representative sample of students (i.e. on a large enough number of students within each classes). For this reason, we elected to use a larger sample of 2541 students—those who had classroom membership information and results on at least one study variable, of 89 classes with between 11 and 34 students (M = 29, SD = 4, with only 3 classes including less than 20 students). These students all attended elementary schools in an inland southern California school district and were tested in the 2005-2006 academic year (T1) and again one year later in 2006-2007 (T2). Here, as in Fast et al. (2010), we focus on data collected in the second year, using T1 data as controls. In this sample used here: (i) 48% were in fifth grade in the second year of the study and 52% were in sixth grade; (ii) 49% were males and 51% were females; (iii) 68% received free or

reduced-price (subsidized) lunches due to coming from low income families (most schools were located in low- to middle income neighbourhoods).

The main variables used in the present study were assessed with subscales from Karabenick and Maehr's (2004, 2007) Student Motivation Questionnaire (SMQ), which was administered in class by teachers who read aloud the questions after providing some examples. Teachers were asked to trade rooms with colleagues so as not to administer the questionnaire to their own students. Four items from the SMQ were used to assess students' math self-efficacy (e.g. "I'm sure that I can learn everything taught in math"), four items were used to assess classroom mastery goal structure (e.g. "My teacher thinks it's important to understand our math work, not just memorize it"), four items were used to assess classroom challenge (e.g. "Our math teacher pushes us to take on challenging work"), and three items were used to assess teacher caring (e.g. "Our math teacher cares about how we feel"). Estimates of composite reliability (calculated with Cronbach alpha coefficient) based on individual students ratings (.84, self-efficacy; .62, mastery; .61, challenge; .75, caring) were almost identical to those reported by Fast et al. (2010), although the climate measures were still at the lowest range of acceptability. This reinforces the need to rely on models incorporating a correction for measurement error. However, we note that composite reliability estimates based on non-disaggregated L1 ratings of a naturally L2 construct such as classroom climate are problematic in themselves and may explain this apparent low level of reliability, we come back to this issue later. These items were rated on a five-point Likert scale ranging from not at all true to very true. Math achievement was assessed with the California Standard Test (CST) of mathematics results obtained at the end of both academic years from the district databases. The CST is a 65-item questionnaire where item content is based on California curriculum standards defining the expected knowledge and skills acquisitions associated to specific grade levels. For the grade levels covered in this study, the CST includes questions related to "number sense", "algebra and functions", "measurement and geometry", and "statistics, data analyses, and probability". Total scores are calculated by summing the number items with correct answers, although only the total scores were available for the analyses and were obtained from the district data bases. Similarly, information regarding gender, lunch status and grade level was also obtained from the district data bases. The SMQ was administered in the spring of each academic year, while the CST

was completed at the end of the academic year, allowing for a within-year temporal ordering between SMQ and CST data.

Results

All analyses were conducted with Mplus (version 6.1; Muthén & Muthén, 2010) using the doubly-latent ML-SEM described in greater detail by Marsh et al. (2009; Lüdtke et al., 2008; 2011). Since a main objective of the present study is to illustrate the use of Ml-SEM models, we provide a longer than usual analytical presentation, in which we address the preliminary verifications that need to be routinely conducted before estimating these models, and an extensive discussion of the reasons underlying some of the model specification decisions that we made.

Preliminary Verifications of Statistical Assumptions and Requirements.

Statistical assumptions requirements in the context of ML-SEM models are not that different from traditional assumptions of regular multilevel or SEM analyses. For instance, common assumptions of most multivariate analyses have to do with missing data and multivariate normality. However, doubly latent multilevel models are routinely estimated with the Mplus statistical package based on the robust Maximum Likelihood (MLR) estimator, which has been found to be efficient in the estimation of latent variable models based on non-normally distributed responses and items rated on answer scales including five or more response categories (e.g., Beauducel & Herzberg, 2006; Dolan, 1994; DiStefano, 2002; Muthén & Kaplan, 1985; Rhemtulla, Brosseau-Liard & Savalei, 2010). Similarly, it was possible to use a larger sample than the original Fast et al. (2010) study because we relied on Full Information Maximum Likelihood (FIML)—rather than a quasi-listwise deletion strategy—to handle missing data (Enders, 2010; Little & Rubin, 1987; Schafer, 1997). FIML estimation, especially when used in conjunction with MLR, has been found to result in unbiased parameter estimates under even very high level of missing data (e.g., 50%) under Missing At Random (MAR) assumptions, and even in some cases to violations of this assumption (e.g. Enders, 2001, 2010; Enders & Bandalos, 2001; Graham, 2009; Larsen, 2011; Shin, Davidson, & Long, 2009). MAR assumes that the propensity for missing data on a variable can be related to other variables in the analysis, but not to levels of the variable itself, a situation that has previously been argued to be the norm in school-based studies conducted in countries with mandatory education, where the most

common source of attrition is student mobility (e.g., Baraldi & Enders, 2010; Graham, Hofer, Donaldson, MacKinnon, & Schafer, 1997; Enders, Dietz, Montague, & Dixon, 2006). FIML is generally recognized to perform equivalently, or even better in some cases in the context of multilevel analyses (e.g. Larsen, 2011), than more computationally intensive multiple imputation procedures (e.g. Enders, 2010; Graham, 2009).

In a series of simulation studies, Lüdtke et al. (2008, 2011) found that sample size, both in terms of the total number of L2 units considered in the study and in terms of L1 members of each of these L2 units, is probably the most important requirement of doubly latent multilevel models. They note that without at least 50, but ideally 100, L2 units (i.e., classrooms) including at least 10-15 participants (i.e., students) per unit, these complex models tend to present higher than acceptable rates of nonconvergence and estimation errors. They also note that the representativeness of the samples in each L2 unit (and within-classroom sampling ratio) should not be too low. With a total of 89 classrooms including an average of 29 students per classroom (with only 3 classes including less than 20 students) - close to the average classroom size usually found in the US - these assumptions are all reasonably met in the present study. It is interesting to note that, using the restricted data set initially used by Fast et al. (2010), most of the models estimated in the present study failed to converge on proper solutions, which can be attributed to their lower total sample size (n = 1163) as well as to the lower within class sample size (88 classes, but with an average of 13 students per class, SD = 6, range 2-28, with only 17 classes including more than 20 students).

Other assumptions that should routinely be checked have to do with multicollinearity at L1 and L2, and whether the a priori measurement models holds at both levels. First, it should be noted that, in order to facilitate interpretation of the results and to reduce non-essential multicollinearity, all variables were standardized (M = 0, SD = 1). We also directly examined potential problems of multicollinearity in individual students' responses. Although examination of the correlation matrix revealed no potential problems of multicollinearity in individual ratings of the constructs used in the present study, we still performed systematic tests which showed that both Tolerance (>1) and VIF (< 2) values remained reasonably low. This observation, together with the verification that the a priori factor model fits the data well at both L1 and L2 is particularly important in doubly latent multilevel

models since they decompose the total variance of the indicators into separate L1 and L2 components. Indeed, even if responses have a well-defined factor structure and no multicollinearity at L1, there is no guarantee that this will transpose at L2. The preliminary multilevel confirmatory factor analysis models used in this study will provide a good illustration of this phenomenon.

Since multilevel models are based on the disaggregation of the total variance into L1 and L2 components, there needs to be a significant amount of variability at L2 to justify investing the efforts required to properly disaggregate L1 and L2 components of these effects. Similarly, another requirement of doubly latent multilevel models has to do with the reliability of the constructs at L1 and L2, and of the class aggregates. The reliability of the latent constructs themselves is generally assessed using scale score composite reliability indicators and verifying that the a priori measurement models fit the data well(we come back to these issues later) in order to confirm that there is a sufficient level of agreement between the items used to estimate the latent constructs. However, these models are called doubly latent since they also rely on a latent aggregation of students' responses to form classroom aggregates. In this perspective, there should also be a good level of agreement, or reliability, between the ratings provided by the students forming each classroom. Lüdtke et al. (2008, 2011) noted that this requirement interacts with sample size requirements so that larger samples are required to achieve proper estimation when reliability is low.

The agreement between any pair of students within the same class (analogous to the average correlation among items in test scores) is assessed with the intraclass correlation coefficient (ICC1), sometimes referred to as the single-rater reliability. The ICC1 also reflects the proportion of the total variance that occurs at the second level (L2) of analysis and is calculated as $\frac{\tau_x^2}{\tau_x^2 + \sigma_x^2}$ where τ_x^2 is the between-group variance and σ_x^2 is the within-group variance. Following Bliese (2000), Raudenbush and Bryk (2002) and others, we distinguish between ICC1 as the average agreement between pairs of students within the same class (i.e., the proportion of the total variance occurring at the classroom level) and ICC2 as the reliability of the group average (analogous to the reliability of a factor based on multiple items; Lüdtke et al., 2008, 2011; Marsh et al., 2009, 2012). ICC2 is computed as

$\frac{\tau_x^2}{\tau_x^2 + (\sigma_x^2 / n_j)}$ where n_j is the average size of the groups. ICC1 values are seldom larger then .3, but should ideally be close to or higher than .1 (Lüdtke et al., 2008, 2011) and reflect the proportion of variance occurring at L2. Hedges and Hedberg (2007) report that values for ICC1 on a variety of studies of school performance in American schools are typically in the .10 to .25 range, whilst values for non-performance measures are typically somewhat more modest. ICC2 values depend on the ICC1 and the number of students within each classroom so that satisfactory values (e.g., .7 to .8 or higher) can be obtained as long as the number of students is adequate. ICC2 values are interpreted in line with other reliability measures, i.e. exceeding or close to 0.8 (e.g., Marsh et al., 2012), and are akin to classical reliability estimates but based on agreement between students rather than agreement among items. These indices were satisfactory and justify the use of doubly latent multilevel models: math self-efficacy, ICC1 = .083 and ICC2 = .726; math achievement, ICC1 = .101 and ICC2 = .766; mastery climate, ICC1 = .160 and ICC2 = .846; challenge climate, ICC1 = .164 and ICC2 = .850; caring climate, ICC1 = .298 and ICC2 = .925. These values are all satisfactory and justify the use of doubly latent multilevel models.

Main Model Specification

The main model assessed in the present study is presented in Figure 2 (see the online supplemental materials: Appendix 1 for the Mplus code and Appendix 2 for additional statistical considerations). Initially, separate models were evaluated for each of the three classroom climate constructs. Then, we explored alternative models that included all three facets of classroom climate. However, due to severe multicollinearity problems at L2 (L2 climate variables were correlated at r = .77 to .92 in preliminary multilevel CFA models), these models failed to converge to proper solutions. Thus, classroom climate was specifically modeled as a higher-order L2 factor based on the three L2 climate factors (see Figure 2). As previously mentioned, we consider classroom climate as a purely L2 construct and had no rationale supporting the examination of the relations between the disaggregated residual L1 ratings of classroom climate and the other variables. However, in order to achieve proper disaggregation of the L1 and L2 components of students' ratings of classroom climate, to control for

unreliability in the aggregation to L2 of these ratings, and to control for sampling error in the context of a ML-SEM model, classroom climate items needed to be specified at both levels and simply allowed to correlate at L1 while their expected effects are modeled at L2. Given the total number of climate items, it was more parsimonious to specify the same higher-order factor structure at L1 and L2, resulting in more stable estimation given that this model only involves the estimation of correlations between a single higher-order factor (versus 11 items) and the other variables. We note however that there is no need to estimate a similar measurement or path model at both levels in the context of doubly latent multilevel models. Indeed, a strength of these models is to allow the examination of different models or research questions at L1 and L2. In fact, the only case where the same paths need to be estimated at both levels is when a contextual effect is estimated, given the need to properly control for the same relation at L1 in interpreting the L2 relation (see the introduction and Appendix 2 for further discussions of contextual effects).

To ensure comparability of results, we controlled for the effects of the same covariates as Fast et al. (2010; the dotted lines in Figure 2): gender and free lunch, two individual students' characteristics, were included as L1 controls, and grade level, a characteristic common to all students within a class, was included as L2 control. In the case of grade level, it proved impossible to control this variable at L1 since there was no variability in grade level at L1. In the case of gender and free lunch, they also only had a very low level of variability at L2 and we had no reason to expect that class proportion of females or of students receiving free lunches would affect the observed relationships. This assumption is also supported by the fact that the classrooms were highly similar to one another in terms of gender and free-lunch compositions, resulting in very low levels of L2 variability on these variables (for instance, the ICC1 for gender was only of .004), precluding proper L2 analyses.

Another difference between the L1 and L2 models is illustrated by the fact that T1 levels of academic achievement and self-efficacy where only modeled at L1. ML-SEM models can only take into account one indicator of nesting into L2 units (but see Beretvas, 2011 for extensions). Here, we took into account students' membership into their classrooms at T2. Since classroom membership change over time, at L2, T1 levels of math achievement or math self-efficacy would represent the average previous level of math achievement or math self-efficacy of the students forming each class

when they were in their previous classes, the year before. Although arguments could likely be made to justify studying L2 average levels of previous achievement or efficacy of the students forming a class, we elect to remain consistent with Fast et al.' (2010) model in which T1 levels on these constructs were only included as controls of previous individual-students levels of math achievement or math self-efficacy. We only used these variables as controls at the individual-student L1 level. In fact, it also made no sense to control for these previous levels, since, they also presented a very low level of L2 variability. For instance, the ICC1 (.016) and ICC2 (.316) for T1 math self-efficacy were clearly too low to justify studying this construct at L2. We also included, in the L1 part of the model, an extension from Fast et al.'s (2010) model. Fast et al. (2010) only allowed T1 math self-efficacy to influence T2 math self-efficacy and T1 math achievement to influence T2 math achievement (see Figure 1). Consistently with current knowledge on relations between self-beliefs and performance (e.g. Marsh & Craven, 2006; Marsh & O'Mara, 2008; Marsh, Trautwein, Lüdtke, Köller & Baumert, 2005; Valentine, DuBois, & Cooper, 2004), we allowed T1 achievement and self-efficacy to influence both T2 self-efficacy and achievement.

Regarding the measurement models, in accordance with recommendations for longitudinal research (Marsh & Hau, 1996; Jöreskog, 1979), correlated uniquenesses were posited a priori between matching indicators of math self-efficacy at Time 1 and 2. Failure to do so has been shown to result in positively biased estimates of stability and distorted parameter estimates (Marsh & Hau, 1996). Because Fast et al. (2010) had not considered multiple indicators in their main analysis, they were not able to control for this source of bias in their main analyses. Since T1 levels of math self-efficacy were not modelled at L2, there was no need to include these correlated uniquenesses at L2.

Finally, as in the Fast et al. (2010) study, we postulated fully mediated models in which classroom climate predicts math self-efficacy, which in turn predicts math achievement, but we modeled classroom climate as an L2 construct. We also conducted tests of partial mediation (the dashed line in Figure 2) where classroom climate variables also directly predicted math achievement.

Goodness of fit was assessed with the Root Mean Square Error of Approximation (RMSEA), the Tucker-Lewis Index (TLI), and the Comparative Fit Index (CFI), as operationalized in Mplus in association with the MLR estimator (Muthén & Muthén, 2010) as well as the robust $\chi 2$ test statistic

and inspection of parameter estimates. Traditional cut-off scores indicative respectively of excellent and adequate fit to the data were used: (i) CFI and TLI ≥ .95 and ≥ .90; (ii) RMSEA ≤ .06 and ≤ .08. The relative fit of different nested models was also compared on these indices (Marsh, 2007b; Marsh, et al., 2009), and information criteria were also used to this end (Akaïke: AIC; Bayesian: BIC; and sample size adjusted BIC: ABIC), with the model with the lowest providing better fit to the data. We note here that unstandardized and standardized coefficients obtained in the context of ML-SEM are interpreted exactly as they are in the context of classical regression or SEM studies, where unstandardized coefficients represent the amount of change in the outcome that can be expected from a unit change in the predictors, and standardized coefficients express the same but in standard deviation units. We also provide effect size indicators that can be interpreted according to Cohen's guidelines (Cohen, 1988). For instance, Cohen (1988) suggests that values over .10, .30, and .50 approximately reflect small, moderate and large effect sizes in the context of regression parameters. Details on how to obtain properly standardized effects and effects sizes are presented in Appendices 1 and 2. Nevertheless, all of such guidelines (i.e. effects sizes and model fit) should be taken as rough guidelines, not golden rules (Marsh, Hau & Grayson, 2005; Marsh, Hau, & Wen, 2004).

Multilevel CFA Models and Invariance over Time and Level.

We first estimated multilevel CFA models in order to ensure that the proposed measurement model fitted the data well and to test whether factor loadings were invariant across levels. Invariance of the factor loadings is not a necessary pre-requisite for doubly-latent multivariate models assessing climate constructs, but still has many advantages. First, invariance of the factor loadings equates the latent factor metric across levels, making them more directly comparable (Lüdtke et al., 2011; Mehta & Neale, 2005). Second, it also provides important information as to whether the estimated constructs are the same at both levels, which is a necessary condition for the estimation of contextual effects given the need to properly control for the same relation at L1 in interpreting the L2 relations (see Appendices 1 and 2, and Lüdtke et al., 2011; Marsh et al., 2009, 2012). Finally, it reduces the complexity of the model, making it more parsimonious and allowing information from L1 to be used in the estimation of the more unstable L2 model. Lüdtke et al. (2011) simulation study even shows that, for this reason, specifying invariant loadings across levels tends to produce more accurate

parameter estimates, even when this specification is erroneous in the population model.

Similar invariance constraints were also imposed on the factor loadings for math self-efficacy across both time points at L1 (math self-efficacy was the only multi-item construct measured at both time points Time 1 and Time 2) to ensure that the meaning of this construct did not shift over time. The fit indices for these models (Table 1) demonstrate that they provided an adequate fit to the data. Importantly, the constructs were reasonably invariant across time and levels—as indicated by the overall absence of decrement in fit indices (and even a slight improvement in the TLI and RMSEA that take into account parsimony) and by the decrease in information criterion indexes. The results from all CFA models based on either single climate factors or on the second-order climate factor led to the same conclusions. Furthermore, ICC1 (.199) and ICC2 (.878) coefficients for the higher-order climate construct were fully satisfactory and of the same magnitude as for the first-order climate constructs. The specific results for the final multilevel second-order CFA model (where climate indicators were represented by three latent constructs themselves associated with a higher-order factor) are reported in Appendix 3.

These results show that all loadings are significant and mostly substantial, confirming the adequacy of the measurement models (see supplemental materials, Appendix 3). These results also confirm the appropriateness of the controls that are significantly, albeit weakly, related to math achievement at both time points (i.e., free lunch : T1 r = -.106, s.e. = .028, $p \leq .05$; T2 r -.113, s.e. = .027, $p \leq .05$), as well as to inter-individual differences in L1 climate ratings (i.e., gender , with r = .122, s.e. = .027, $p \leq .05$). Similarly, at L2, grade level is significantly and moderately related to climate (r = -.268, s.e. = .113, $p \leq .05$) and achievement (r = -.229, s.e. = .105, $p \leq .05$). The results show that, as expected, math self-efficacy and achievement tend to be reciprocally related longitudinally at L1 (T1 efficacy to T2 efficacy: r = .461, s.e. = .031, $p \leq .05$; T1 efficacy to T2 achievement: r = .271, s.e. = .026, $p \leq .05$; T1 achievement to T2 efficacy: r = .294, s.e. = .025, $p \leq .05$; T1 achievement to T2 achievement: r = .800, s.e. = .012, $p \leq .05$). Of particular importance, the higher-order climate factor has only small to moderate correlations with other constructs at L1 ($|r|$ = .001 to .467, M = .139, SD = .176) but is substantially related to Time 2 math self-efficacy at L2 (r = .688, s.e. = .098, $p \leq .05$; L2 correlations varying between $|r|$ = .260 and .688, M = .405, SD = .245).

This confirms the appropriateness of treating climate as an L2 construct.

Composite reliability coefficients were calculated from the parameters estimates from this multilevel CFA model in order to estimate the amount of measurement error present in the estimation of the different latent constructs at both L1 and L2. To this end, we used McDonald's (1970) $\omega = (\Sigma|\lambda i|)^2 / ([\Sigma|\lambda i|]^2 + \Sigma\delta ii)$ where λi are the factor loadings and δii, the error variances. Compared with traditional scale score reliability estimates (e.g., alpha; see Sijtsma, 2009), ω has the advantage of taking into account the strength of association between items and constructs (λi) as well as item-specific measurement errors (δii) and to be applicable to the estimates obtained at both L1 and L2 based on level-specific variance-covariance matrices. These coefficients are interpreted as any other composite reliability coefficients and are fully satisfactory for math self-efficacy ratings at both L1 (ω = .79 at T1 and .84 at T2) and L2 (ω =.95 at T2), as well as for the higher-order climate factor (ω = .88 at L1 and .91 at L2). However, confirming the results from the previously reported coefficient alpha estimates for individual ratings, the estimated composite reliability coefficients were at the lowest range of acceptability for the specific climate subscales at L1 (ω = .57 for challenge, .71 for caring and .59 for mastery), confirming the importance of using models that incorporate a control for measurement error. More precisely, a latent variable model will estimate the higher-order factor from the specific climate factors, themselves estimated net of measurement errors. It is thus not surprising that the higher-order climate factor itself is highly reliable, since it is estimated from lower-order constructs from which measurement error has been partialled out. However, these composite reliability coefficients for the specific climate factors were substantially higher and fully satisfactory at L2 (ω = .78 for challenge, .88 for caring and .89 for mastery). In other words, doubly latent models also partial out another source of unreliability when estimating L2 constructs: agreement between students forming each classroom. This form of reliability is estimated from the previously reported ICC2. This suggests that, although when individual students are asked to rate their classroom climates, they provide highly unreliable information. However, when the agreement between students forming a class (similar to the agreement between the items forming a construct) is considered in the aggregation of these ratings to L2, average class ratings of classroom climate are found to be highly reliable. This clearly confirms our decision to consider climate as a L2 construct

Main Results: Climate Effects on Math Self-Efficacy and Achievement

Our main objective was to investigate the effects of classroom climate on math self-efficacy and achievement using properly specified ML-SEM models. We first estimated three separate models based on each of the three climate factors and then a final model, including a single higher-order climate factor. For all of these models, in order to investigate the proposed mediation hypothesis more fully, two alternative specifications were considered. In the first, climate effects on math achievement were posited to be fully mediated by math self-efficacy (i.e. no direct effects of climate on achievement were specified in the model). In the second, these effects were only partially mediated (i.e. direct effects were posited as well as indirect effects: see the dashed path in Figure 2).

Comparisons of these alternative models in terms of fit (see Table 1) yielded identical conclusions, favouring the fully mediated model for all climate dimensions, as well as for the higher-order factor model. Interestingly, moving from the multilevel CFA to the fully mediated MLSEM models resulted in almost no decrease in fit and in lower information criteria (e.g., for the model with the higher-order factor: CFI = .960 for the CFA *vs* .960 for the ML-SEM; TLI = .953 *vs* .953; RMSEA = .024 vs .024; AIC = 131494 vs 131363; BIC = 132177 vs 132029; ABIC = 131805 vs 131667). This suggests that the more parsimonious ML-SEM models (i.e., were a limited number of predictive paths replace freely estimated correlations between all constructs) provided an adequate representation of relations among the variables. Adding partial mediation to these models did not affect the fit of the model (e.g., for the model with the higher-order factor, the CFI, TLI and RMSEA remained the same), but resulted in increases in information criteria values (e.g. for the model with the higher-order factor, the AIC increased from 131363 to 131365, the BIC increased from 132029 to 132036, and the ABIC increased from 131667 to 131671), suggesting that these models did not represent an improvement over the fully mediated models. Indeed, inspection of the partially mediated models revealed that the added path was never statistically significant ($p < .05$). Thus, fully mediated models were retained. The main results from the final higher-order ML-SEM model are reported in Table 2. The specific climate effects from the models including a single climate dimension are also reported in greyscale.

Examination of these results shows significant L2 effects of climate dimensions on Time 2 math self-efficacy, albeit of moderate magnitude. Interestingly, standardized effects and effects sizes

were of a similar magnitude for each of the climate dimensions considered separately (standardized estimates = .137 to .173, with effect sizes = .284 to .357), as well as for the higher-order climate factor (standardized estimate =.172, with effect size =.357). Moreover, when the percentage of the variance of L2 math self-efficacy is explained by the model (R^2, routinely provided as part of Mplus outputs), it remains quite similar when the higher order climate factor is considered (R^2 = .490), when the three climate dimensions are entered together (R^2 = .508; as previously noted however, this model was not fully proper, due to multicollinearity), and when the three climate dimensions are entered separately (R^2 = .324 for caring climate, .452 for challenge climate, and .493 for mastery climate).This suggests that the L2 effects represent generic effects of classroom climate quality rather than the specific effects of climate dimensions. The contextual effect of L2 classroom levels of math self-efficacy on math achievement is also significant, and of a moderate-small magnitude (standardized estimate =.073, with effect size =.154) and of a comparable magnitude to the L1 effect of individual levels of math self efficacy on math achievement (standardized estimate =.086, with effect size =.181). Regarding the mediation hypothesis, the L2 indirect effect of classroom climate on math achievement as fully mediated by math self-efficacy is also significant (indirect effect = .359; SE = .158; $p \leq .05$; 95% confidence interval = .049 to .668).

Finally, for comparison purposes, Fast et al. (2010) results, as well as results from the same model estimated on the larger sample with FIML estimation, are reported in Table 3. The results presented in this table here are very similar to those reported by Fast et al. (2010) in showing complete mediation through which classroom climate influences students' math self-efficacy, which in turn influences math achievement. However, in comparison with our main results, these results assume that all effects occur at L1, whereas our main results show that these relations are in fact located at L2.

Discussion

Classroom Climate Effects.

Substantively, our results demonstrate that classroom climate does predict classroom levels of math self-efficacy and achievement. This indicates that classroom levels on these variables do depend on the quality of classroom climate in the target year. Although longitudinal correlational studies like the present investigation preclude clear causal conclusions (also see Marsh et al., 2012), the results

suggest that climate interventions should target whole classes, and have the potential to improve both self-beliefs and achievement at the level of the classroom as a whole, thus resulting in improvements for individual students within each classroom.

Our results may appear to be quite similar to those reported by Fast et al. (2010) who also show complete mediation through which classroom climate perceptions influence individual students' levels of math self-efficacy, which in turn influences their levels of math achievement. However, our results differ from theirs in a substantively important manner. Indeed, our results show that these effects are classroom climate effects and clearly located at the classroom level whereas disaggregated residual L1 ratings by individual students presented a very low level of reliability. This finding is important and reinforces the inappropriateness of making interpretations based on L1 student ratings of classroom climate, as in Fast et al.'s (2010) study and many other classroom climate studies.

It is relevant to note that the regression parameter estimates from Fast et al.'s (2010) study are roughly similar in size and direction to those observed in our study. The explanation for the similarity of the results is the complicated confounding of L1 and L2 effects of climate ratings in the Fast et al. analysis that are disentangled in our ML-SEM model (see Marsh, et al., 2012, for further discussion of this issue). Indeed, the L1 effects in the Fast et al. study (what they called 'perceived climate effects') that were not controlled for L2 effects actually represented—at least in part—real L2 classroom climate effects rather than—or in addition to—residual 'perceptive' effects at the level of individual students. Hence, the effects observed by Fast et al. could not solely be attributed to individual students' perceptions as implicit in their study, but rather were confounded with the shared effects of students within each classroom. Because *classroom* climate is, by definition, inherently a classroom level construct, this distinction makes a fundamental difference in the interpretation of the results. Furthermore, even if there were substantively meaningful effects of the residualized L1 climate ratings, these could not be appropriately identified in single-level models that confound the effects of L1 individual students' perceptions of climate and the L2 effects of climate. Thus, even if the objective of a study is to study the effects inter-individual differences in perceptions of the classroom climate, these perceptions should be appropriately disentangled from classroom-level effects in the context of appropriate multilevel models. Particularly if there were effects of residualized L1 climate

ratings, these might be inappropriately interpreted to reflect classroom climate in a single-level model. In summary, L1 ratings of classroom climate in single-level models should not be used to represent either L2 classroom climate constructs or individual differences in student perceptions at L1.

Differentiation among classroom climate constructs.

Our results also have implications for differentiation among different classroom climate constructs. In our study (and in the original study by Fast et al., 2010), individual students' perceptions of their classroom climate could be differentiated at L1 into the three a priori components of classroom climate factors (caring, challenge and mastery). Based on L1 responses, Fast et al. (2010) reported correlations among the three climate factors as .34 to .50 (see their Table 1). They discussed these "high" correlations as a possible limitation of their study, but also suggested that 'These constructs are often highly correlated with one another because there is a common denominator underlying them: a teaching disposition that anticipates and is responsive to student learning needs' (p. 738). However, when classroom themselves became the unit of observation (L2) in our ML-SEM model and adjusted for measurement and sampling error, these correlations were much higher (.77 to .92) and were so highly correlated that they could not be distinguished. We resolved this issue by positing a single higher-order indicator of classroom climate quality that fitted the data as well as the three L2 classroom climate factors considered separately. This suggests that, consistent with suggestions by Fast et al. (2010), the separate climate factors considered here are apparently characterized on a single global climate continuum reflecting a "teacher disposition that anticipates and his responsive to students' learning needs". Although this observation probably does not generalize to all possible facets of classroom climate (e.g. Adelman & Taylor, 2002; Fraser, 1998; Trickett & Moos, 1995), it highlights the importance of treating classroom climate as an L2 construct in appropriate multilevel models incorporating control for measurement error. Because factor analysis research on classroom climate responses has relied so heavily on single student-level factor analyses, it is important to determine when the different climate factors can be differentiated at the classroom level using multilevel CFAs and SEMs. Thus, these results may have wide psychometric relevance for future studies of classroom or school climates. In particular, they show that even thought individual students appear to be able to differentiate various facets of climate, these distinctions are plagued by a

substantial amount of measurement error and might not generalize to the classroom level that is critical for climate studies. If the present results generalize to other classroom or school climate constructs, they would indicate the need for a critical re-appraisal of previous psychometric work on the measurement of L2 climate constructs.

Alternative Approaches for Smaller Samples.

Clearly, doubly latent ML-SEM models are a large sample analytical procedure and work is still needed to establish best practices when sample sizes at L1 or L2 are modest (see Lüdtke et al., 2008, 2011; Marsh et al., 2009). Indeed, in such cases, the doubly latent ML-SEM advocated here may simply fail to converge to a proper solution (as this was the case for the restricted sample used by Fast et al., 2010), or yield unstable or improper parameter estimates. In this cases, Lüdtke et al. (2008, 2011; for related discussion and details on how to estimate these alternative models, see Marsh et al., 2009) recommend the use and comparison of results based on partial correction models involving only: (a) latent aggregation of L1 manifest variables to form L2 constructs, thus controlling only for sampling error; (b) latent measurement models at L1 and L2 but manifest aggregation to L2, thus controlling only for measurement error; or even (c) doubly manifest models including no controls for measurement or sampling errors. Lüdtke et al. (2008, 2011) simulation results show that when sample sizes are small the relatively small amounts of biases introduced by these procedures tend to be less troublesome than relying on doubly latent ML-SEM models and converging on unstable or improper parameters estimates. The extreme case of having an insufficient number of L2 units to conduct multilevel analyses was recently addressed by Morin, Maïano, Marsh, Nagengast & Janosz (in press), who recommended using single-level analyses based on group-mean centring of the variables in order to obtain appropriately corrected estimates of the L1 model parameters.

Conclusion

Methodologically, our study illustrates a number of desirable features of climate studies and thus, provides guidance for future research. First, it is critical that climate effects are based on either true L2 (classroom, schools, etc.) measures or appropriate aggregates of L1 measures. Second, in appropriately designed measures of classroom climate, the referent should be the classroom, not the individual student. The L1 ratings of climate are important in terms of estimating agreement among

students within the class and forming the L2 aggregates, but the L1 ratings do not represent classroom climate and usually have no substantive meaning in relation to the interpretation of true L2 climate effects (e.g. Cronbach, 1976). If applied researchers offer substantive interpretations of L1 residual climate ratings they should provide a clear theoretical rationale for doing so, and statistical evidence that the L1 residuals from appropriate multilevel models are meaningfully associated with other constructs in conformity with this rationale. However, these residuals should not be interpreted as reflecting meaningful individual characteristics but rather, individual differences from an average perception of classroom climate. Furthermore, strong correlations between residual L1 climate ratings (controlling for L2 classroom climate) and other constructs, might call into question the construct validity of the classroom climate construct itself, or lead to alternative interpretations of the meaning of this construct. If researchers want to evaluate individual difference constructs, then the referent should be the individual student, not the classroom. If researchers want to consider individual students' characteristics and classroom climate simultaneously, then they should include parallel sets of items—one with the individual student as the referent and one with the classroom or teacher as the referent (see, e.g. Papaioannou et al., 2004).

To date, emerging approaches in applied educational studies of classroom climate have relied on either single level latent variable methodologies, or multilevel models based on manifest variables. However, rarely have climate studies combined both approaches—latent variable multilevel models. This is unfortunate, as most educational research is based on constructs that cannot be measured without measurement error and is inherently multilevel, so that it is necessary to disaggregate L1 and L2 effects and to control for measurement and sampling error. Fortunately, recent statistical advances allow for the integration of both approaches and for their implementation in easy to use commercially available packages. The doubly latent ML-SEM model presented here represents an advance that has broad relevance for the fields of educational, psychological, sociological and organizational research, and for the social sciences more generally. Because this approach has not yet been widely applied, much work is still needed to establish best practice and appropriate limitations (see Lüdtke et al., 2008, 2011). Nevertheless, this new framework offers exciting new possibilities for applied researchers in educational psychology and in the social sciences more generally.

References

Adelman, H.S. & Taylor, L. (2002). Classroom climate. In S. W. Lee, P. A. Lowe, & E. Robinson (Eds.), *Encyclopedia of School Psychology* (pp. 304–312). Thousand Oaks, CA: Sage.

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology, 48,* 5-37.

Beauducel, A., & Herzberg, P. Y. (2006). On the Performance of Maximum Likelihood Versus Means and Variance Adjusted Weighted Least Squares Estimation in CFA. *Structural Equation Modeling, 13,* 186–203.

Beretvas, S. N. (2011). Cross-classified and multiple-membership models. In J. J. Hox & J. K. Roberts (Eds), *Handbook of Advanced Multilevel Modeling* (pp. 313-334). New York, NY: Routledge.

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. Kozlowski (Eds.), *Multilevel Theory, Research, and Methods in Organizations* (pp. 349-381). San Francisco, CA: Jossey-Bass.

Ciani, K. D., Middleton, M. J., Summers, J. J., & Sheldon, K. M. (2010). Buffering against performance classroom goal structures: The importance of autonomy support and classroom community. *Contemporary Educational Psychology, 35,* 88–99.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Mahwah, NJ: Erlbaum.

Cronbach, L. J. (1976). *Research on classrooms and schools: Formulations of questions, design and analysis*. Stanford, CA: Stanford Evaluation Consortium.

DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling, 9,* 327–346.

Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5 and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology, 47,* 309–326.

Enders, C. K. (2001). The impact of nonnormality on full information maximum-likelihood estimation for structural equation models with missing data. *Psychological Methods, 6,* 352-370.

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

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.

Enders, C. K., Dietz, S., Montague, M., & Dixon, J. (2006). Modern alternatives for dealing with missing data in special education research. In T. E. Scruggs & M. A. Mastropieri (Eds.), Advances in learning and behavioral disorders, Vol. 19 (pp. 101–130). New York: Elsevier.

Enders, C. K. & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods, 12*, 121–138.

Fast, L. A., Lewis, J. L., Bryant, M. J., Bocian, K. A., Cardullo, R. A., Rettig, M., & Hammond, K. A. (2010). Does math self-efficacy mediate the effect of the perceived classroom environment in standardized math test performance? *Journal of Educational Psychology*, *102*, 729–740.

Fraser, B. J. (1998). Classroom environment instruments: Development, validity, and applications, *Learning Environments Research, 1*, 7–33.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*, 549-576.

Graham, J. W., Hofer, S. M., Donaldson, S. I., MacKinnon, D. P., & Schafer, J. L. (1997). Analysis with missing data in prevention research. In K. J. Bryant, M. Windle & S. G. West (Eds.), The science of prevention: Methodological advances from alcohol and substance abuse research (pp. 325−366). Washington, DC: American Psychological Association.

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.

Jöreskog, K.G. (1979). Statistical models and methods for the analysis of longitudinal data. In K.G. Jöreskog & D. Sörbom (Eds.), *Advances in Factor Analysis and Structural Equation Models.* Cambridge, MA: Abt Books.

Karabenick, S. A., & Maehr, M. L. (2004). MSP-Motivation Assessment Program: First-year report to the National Science Foundation. Retrieved from http://ma.mspnet.org/media/data/mspmap.pdf?media_000000005860.pdf

Karabenick, S. A., & Maehr, M. L. (2007). MSP-Motivation Assessment Program: Final report to the National Science Foundation. Retrieved from http://ma.mspnet.org/media/data/MSP-

MAP_Final_Report.pdf?media_000000006004.pdf

Larsen, R. (2011). Missing data imputation versus full information maximum likelihood with second level dependencies. *Structural Equation Modeling, 18,* 649-662.

Lau, S., & Nie, Y. (2008). Interplay between personal goals and classroom goal structures in predicting student outcomes: A multilevel analysis of person-context interactions. *Journal of Educational Psychology, 100,* 15-29.

Little, R. J. A. & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.

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

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.

Marsh, H. W. (2007a). *Self-concept theory, measurement and research into practice: The role of self concept in educational psychology—25th Vernon-Wall lecture series.* London, UK: British Psychological Society.

Marsh, H. W. (2007b). Application of confirmatory factor analysis and structural equation modeling in sport and exercise psychology. In G. Tenenbaum & R. C. Eklund. (Eds), *Handbook of sport psychology, 3rd edition* (pp. 774–798). Hoboken, NJ: Wiley.

Marsh, H. W., & Craven, R. G. (2006). Reciprocal effects of self-concept and performance from a multidimensional perspective: Beyond seductive pleasure and unidimensional perspectives. *Perspectives on Psychological Science, 1,* 133–163.

Marsh, H. W., & Hau, K-T. (1996). Assessing goodness of fit: Is parsimony always desirable? *Journal of Experimental Education, 64,* 364–390.

Marsh, H. W., & Hau, K-T. (2007). Applications of latent-variable models in Educal Psych: The need

for methodological-substantive synergies. *Contemporary Educational Psychology, 32*, 151-171.

Marsh, H. W., Hau, K-T., & Grayson, D. (2005). Goodness of Fit Evaluation in Structural Equation Modeling. In A. Maydeu-Olivares & J. McCardle (Eds.), *Psychometrics. A Festschrift to Roderick P. McDonald*. Hillsdale, NJ: Erlbaum.

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 overgeneralising Hu & Bentler's (1999) findings. *Structural Equation Modeling*, *11*, 320–341.

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J. S., Abduljabbar, A. S. & Köller, O. (2012). Classroom climate effects: Methodological issues in the evaluation of group-level effects. Educational Psychologist? *Educational Psychologist, 47*, 106-124.

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.

Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O. & Baumert, J. (2005) Academic self-concept, interest, grades and standardized test scores: Reciprocal effects models of causal ordering. *Child Development*, *76*, 297–416.

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

Marsh, H. W., Seaton, M., Kuyper, H., Dumas, F., Huguet, P., Régner, I., Buunk, A. P., Monteil, J.-M., & Gibbons, F. X. (2010) Phantom Behavioral Assimilation Effects: Systematic Biases in Social Comparison Choice Studies. *Journal of Personality*, *78*, 671–710.

McDonald, R.P. (1970). Theoretical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis. *British Journal of Mathematical & Statistical Psychology, 23*, 1-21.

Metha, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods, 10*, 259–284.

Midgley, C., Maehr, M., Hruda, L., Anderman, E., Anderman, L., Freeman, K., Gheen, M., . . . Urdan,

T. (2000). Manual for the Patterns of Adaptive Learning Scales (PALS). Ann Arbor, MI: University of Michigan.

Miller, A. D. (2006). *Teacher–student relationships in classroom motivation: A critical review of goal structures.* Paper presented at the 2006 meeting of the American Psychological Association.

Miller, A. D., & Murdock, T. B. (2007). Modeling latent true scores to determine the utility of aggregate student perceptions as classroom indicators in HLM: The case of classroom goal structures. *Contemporary Educational Psychology, 32*, 83–104.

Morin, A.J.S., Maïano, C., Marsh, H.W., Nagengast, B., & Janosz, M. (in press). School life and adolescents' self-esteem trajectories. *Child Development*.

Muthén, B. O., & Asparouhov, T. (2011). Beyond multilevel regression modeling: Multilevel analysis in a general latent variable framework. In J. J. Hox & J. K. Roberts (Eds), *Handbook of Advanced Multilevel Modeling* (pp. 15-40). New York, NY: Routledge.

Muthén, B. O., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical & Statistical Psychology, 38*, 171–189.

Muthén, L. K., & Muthén, B. O. (2010). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.

Papaioannou, A., Marsh, H. W., & Theodorakis, Y. (2004). A multilevel approach to motivational climate in physical education and sport settings: An individual or a group level construct? *Journal of Sport and Exercise Psychology, 26*, 90–118.

Preacher, K. J., Zhang, Z., & Zyphur, M. J. (2011). Alternative methods for assessing mediation in multilevel data: The advantages of multilevel SEM. *Structural Equation Modeling, 18,* 161–182.

Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods, 15*, 209–233.

Rabe-Hesketh, S., Skrondal, A. & Pickles, A. (2004). *GLLAMM Manual (Tech report 160)*. U.C. Berkeley Division of Biostatistics Working Paper Series.

Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical linear models (2nd ed.). Thousand Oaks: Sage.

R Development Core Team (2011). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rhemtulla, M., Brosseau-Liard, P., & Savalei, V. (2010). *How many categories is enough to treat data*

as continuous? A comparison of robust continuous and categorical SEM estimation methods under a range of non-ideal situations. Manuscript under review, available at: http://www2.psych.ubc.ca/~mijke/files/HowManyCategories.pdf

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15*, 351–357.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data.* Boca Raton, FL: Chapman & Hall.

Schwartz, S. (1994). The fallacy of the ecological fallacy: The potential misuse of a concept and the consequences. *American Journal of Public Health, 84*, 819–824.

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.

Sijtsma, K. (2009). On the use, misuse, and the very limited usefulness of Cronbach's alpha [Introduction to a special issue]. *Psychometrika, 74*, 107-120.

Skrondal, A., & Laake, P. (2001). Regression among factor scores. *Psychometrika, 66*, 563–576.

Trickett, E. J., & Moos, R. H. (1995). *Classroom Environment Scale Manual, 3rd Edition.* Palo Alto, CA: Consulting Psychologists Press.

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.

Table 1.

Fit Indices for the Multilevel CFA and SEM Models

Model	χ^2 (df) †	AIC	BIC	ABIC	CFI †	TLI	RMSEA
Model including only challenge climate							
CFA, no invariance	366 (115)*	87950	88494	88198	.971	.958	.029
CFA, level invariance	357 (121)*	87942	88450	88174	.972	.962	.026
CFA, total (level + time) invariance	357 (124)*	87938	88429	88162	.973	.964	.027
SEM, full mediation	359 (125)*	87809	88282	88024	.973	.964	.027
SEM, partial mediation	357 (124)*	87809	88288	88027	.973	.964	.027
Model including only caring climate							
CFA, no invariance	253 (93)*	79829	80343	80063	.982	.973	.026
CFA, level invariance	259 (98)*	79827	80311	80048	.982	.974	.025
CFA, total (level + time) invariance	259 (101)*	79822	80290	80035	.982	.975	.025
SEM, full mediation	259 (102)*	79691	80141	79897	.982	.976	.025
SEM, partial mediation	259 (101)*	79693	80149	79901	.982	.975	.025
Model including only mastery climate							
CFA, no invariance	260 (115)*	88036	88580	88284	.983	.975	.022
CFA, level invariance	267 (121)*	88034	88542	88266	.983	.976	.022
CFA, total (level + time) invariance	267 (124)*	88030	88520	88254	.983	.977	.021
SEM, full mediation	359 (125)*	87900	88373	88115	.983	.977	.021
SEM, partial mediation	267 (124)*	87901	88380	88119	.983	.977	.021
All climate variables in a second order model							
CFA, no invariance	832 (320)*	131495	132272	131849	.959	.951	.025
CFA, level invariance	848 (333)*	131498	132199	131817	.959	.952	.025
CFA, total (level + time) invariance	847 (336)*	131494	132177	131805	.960	.953	.024
SEM, full mediation	847 (337)*	131363	132029	131667	.960	.953	.024
SEM, partial mediation	847 (336)*	131365	132036	131671	.960	.953	.024

† The apparent non-monotonicity of some of the χ^2 and CFI index is related to the fact that MLR estimation incorporates scaling correction factors. When these correction factors, as well as rounding, are taken into account, the χ^2 and CFIs are monotonic with model complexity.

Note. * = $p \leq .05$; CFA = confirmatory factor analysis; SEM = structural equation model; χ^2 = chi square test of model fit calculated under the robust maximum likelihood (MLR) estimator; AIC: Akaïke information criterion; BIC = Bayesian information criterion; ABIC = sample-size adjusted BIC; CFI = comparative Fit Index; TLI = Tucker-Lewis index; RMSEA = root mean square error of approximation.

Table 2.

Main Effects from the Final Higher-Order Multilevel SEM Model Presented in Figure 2

	Est. (S.E.)	Std. (S.E.)	ES (S.E.)
L1 effects			
T1 achievement → T2 self-efficacy	.116 (.015)*	.171 (.022)*	.355 (.047)*
T1 achievement → T2 achievement	.722 (.023)*	.719 (.022)*	1.517 (.046)*
T1 self-efficacy → T2 achievement	.021 (.031)	.013 (.019)	.027 (.040)
T1 self-efficacy → T2 self-efficacy	.435 (.033)*	.398 (.035)*	.824 (.072)*
T2 self-efficacy → T2 achievement	.152 (.028)*	.086 (.016)*	.181 (.034)*
L2 effects			
T2 climate → T2 self-efficacy (climate)	.611 (.124)*	.172 (.033)*	.357 (.068)*
Mastery only from the single climate models	.643 (.141)*	.173 (.036)*	.357 (.074)*
Challenge only from the single climate models	.660 (.173)*	.166 (.034)*	.343 (071)*
Caring only from the single climate models	.257 (.069)*	.137 (.035)*	.284 (.072)*
T2 self-efficacy → T2 achievement	.587 (.234)*	.073 (.029)*	.154 (.060)*

Note. * = $p \leq .05$; Est. = unstandardized parameter estimate; S.E.: standard error of the estimate; Std. = standardized parameter estimate; ES: effect size.

Table 3.

Standardized Effects (Standard Errors) from Fast et al. (2010) Model presented in Figure 1

	Fast et al. results (N = 1163)[a]	Fast et al. model (N = 2541)
T2 Self-efficacy → T2 achievement	.09 (.018)*	.13 (.017)*
T2 Mastery climate → T2 achievement	.03 (.022)	.02 (.016)
T2 Challenge climate → T2 achievement	.02 (.019)	-.02 (.014)
T2 Caring climate → T2 achievement	-.04 (.022)	-.04 (.017)*
T2 Mastery climate → T2 self-efficacy	.11 (.031)*	.14 (.034)*
T2 Challenge climate → T2 self-efficacy	.12 (.029)*	.13 (.026)*
T2 Caring climate → T2 self-efficacy	.15 (.031)*	.16 (.023)*

* = $p \leq .05$

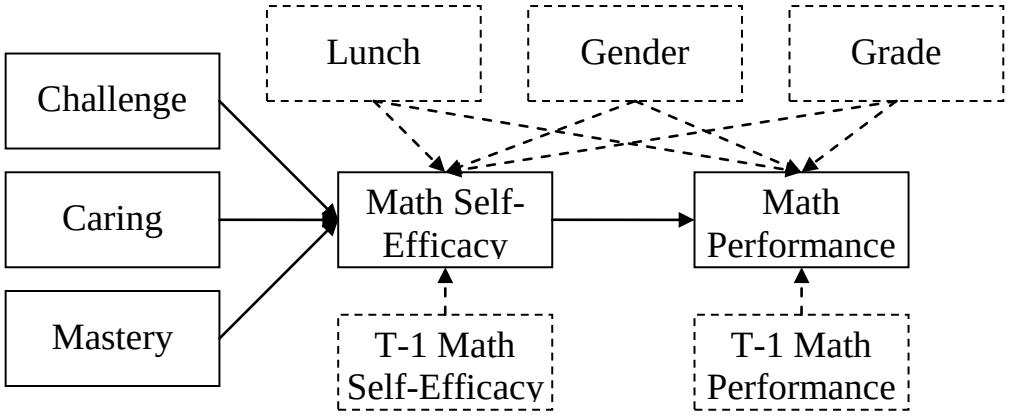

Figure 1. Model tested in the original Fast et al. (2010) study

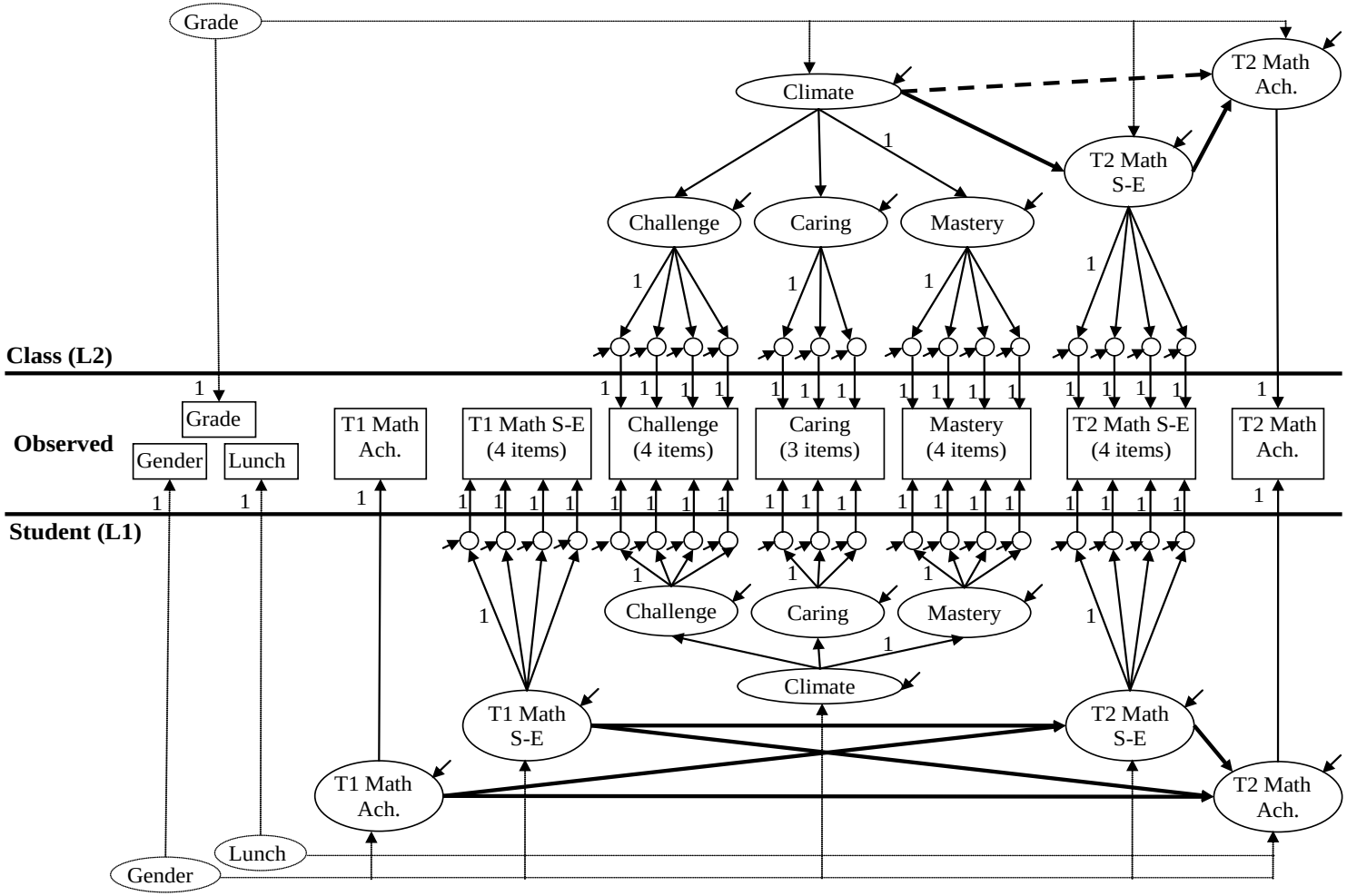

Figure 2. Doubly latent model tested in the current study. *Note.* Dotted lines represent controlled variables; dashed lines refer to optional paths involved in tests of partial mediation; L1 factor correlations between Time 1 achievement and self-efficacy and between climate and Time 1-Time 2 achievement and self-efficacy, as well as correlated uniquenesses between self-efficacy items used at Time 1 and T2 (L1) not shown in the figures.

ONLINE SUPPLEMENTS for

Doubly Latent Multilevel Analyses of Classroom Climate: An illustration

Note: These supplemental appendices will appear on an external website that is hot-linked to the published article.

Appendix 1: Mplus Code for the Final Model

To our knowledge, Mplus is yet the only "user-friendly" statistical package implementing doubly latent multilevel models using non-technical code. However, technically-oriented users could also implement these models using either the R (R Development Core Team, 2011) or GLLAMM (Rabe-Hesketh, Skrondal, & Pickles, 2004) statistical packages.

```
! In Mplus code, the text following "!" symbols on the rest of the line is ignored.
! Each command finish with a ";"

! Optional statement to provide a title to the input.
TITLE: Re-analysis of the JEP data set, two-level model with higher order factor, loading invariance, full
mediation

! Statement to indicate the name and location of the data file
DATA: FILE IS ' C:/DATA.dat';

VARIABLE: ! section off the input to define the nature of the data set
! List, in order of appearance, the variables include in the data set.
NAMES ARE STUDID DID06 SCID06 MA606 CH1306 MA1406 CH1506 EFF1906 MA2306 CH2706
CA2906 EFF3106 CA3206 EFF3506 CH3706 CA4306 EFF4406 MA4806  DID07 SCID07 MA607
CH1307 MA1407 CH1507 EFF1907 MA2307  CH2707 CA2907 EFF3107 CA3207 EFF3507 CH3707
CA4307 EFF4407 MA4807 grad06 CSTRAW06 CSTSCA06 GRADE07 CSTRAW07 CSTSCA07
FRELUNC GENDER  xgrade07 ;

! List the variables to use in the model. Include variables created later with the "DEFINE"
! command at the end.
USEVAR ARE  EFF1906 EFF3106 EFF3506 EFF4406 MA607 CH1307   MA1407 CH1507 EFF1907
MA2307 CH2707 CA2907 EFF3107 CA3207 EFF3507 CH3707 CA4307 EFF4407 MA4807 CSTRAW06
CSTRAW07  GENDER  FRELUNC glev_b;

! Indicate the individual identification code, when there is one, the code use to identify missing data,
! and the code to identify membership into the classes (i.e. the L2 identification code).
IDVARIABLE = STUDID;
MISSING ARE ALL (9999999);
CLUSTER = DID07;

! List of variables that are only used within (in the L1 part of the model)
Within =    GENDER FRELUNC EFF1906 EFF3106 EFF3506 EFF4406 CSTRAW06;

! List of variables that are only used between (in the L2 part of the model)
Between = glev_b;

! Centering the variables that are only used within (in the L1 part of the model)
  CENTERING = GRANDMEAN (GENDER FRELUNC);

Define:  ! This function is used to create new variables from those already in the data set.

! Standardizing all variables except those that are only used at L1 prior to the analysis.
! This is only used to simplify the interpretations and to reduce non-essential multicollinearity.
Standardize MA607 CH1307 EFF1906 EFF3106 EFF3506 EFF4406 CSTRAW06 MA1407 CH1507
EFF1907 MA2307 CH2707 CA2907 EFF3107 CA3207 EFF3507 CH3707 CA4307 EFF4407 MA4807
CSTRAW07;
```

```
! Creating a manifest L2 aggregation of the grade-level control  variable
! since students from the same class all have the same grade-level.
glev_b = CLUSTER_MEAN (xgrade07);

!This is the part were technical specifications are indicated.
ANALYSIS:
TYPE = TWOLEVEL;  ! to indicate a multilevel model
Process = 3;   !To estimate using three processors
ITERATIONS = 10000; ! To increase the number of iterations to help converge on a proper model.
H1ITERATIONS= 10000;  !Increase iterations in estimating the unrestricted model with missing data.


MODEL: !This is the part where the model is specified
%Within%  ! This is for the L1 part of the model.
! Measurement model, where factors are defined with BY statements.
! Parameters followed by the same parameter label in parentheses, e.g. (9) are estimated as invariant.
! Non-invariant parameters can also be labelled for use in conjunction with the model constraints.
! Time 1 efficacy
EF1_W BY EFF1906@1 (9);
EF1_W BY EFF3106 (10);
EF1_W BY EFF3506 (11);
EF1_W BY EFF4406 (12);
! Time 2 efficacy
EF2_w BY EFF1907@1  (9);
EF2_w BY EFF3107 (10);
EF2_w BY EFF3507 (11) ;
EF2_w BY EFF4407 (12);
! correlated uniquenesses across T1-T2 efficacy, correlations are defined with the WITH statement.
EFF1906 WITH EFF1907;
EFF3106 WITH EFF3107;
EFF3506 WITH EFF3507;
EFF4406 WITH EFF4407;
!Time 1 Achievement . Since a single indicator is used, the latent is fixed, with @,
! to be equal to the observed value at L1
ACH1_w BY CSTRAW06@1; CSTRAW06@0;
! Time 2 Achievement
ACH2_w BY CSTRAW07@1; CSTRAW07@0;
! Mastery climate
MA_w BY MA607@1 (1) ;
MA_w BY MA1407 (2) ;
MA_w BY MA2307 (3);
MA_w BY MA4807 (4) ;
! Challenge climate
CH_w BY CH1307@1   (5);
CH_w BY CH1507 (6);
CH_w BY CH2707    (7);
CH_w BY CH3707  (8);
! Caring Climate
CA_w BY  CA2907@1  (13);
CA_w BY  CA3207 (14);
CA_w BY CA4307  (15);

! Second order climate
clim_w BY MA_w@1 (16);
clim_w BY CH_w (17);
clim_w BY CA_w (18);
```

```
! predictive paths (defined with ON).
EF2_w ON EF1_W (b_e2e1_w);
ACH2_w ON ACH1_w (b_a2a1_w);
EF2_w ON ACH1_w (b_e2a1_w);
ACH2_w ON EF1_W (b_a2e1_w);
ACH2_w ON EF2_W (b_a2e2_w);

! Correlating out L1 climate. L1 climate indicators need to be part of the model to properly
! aggregate them at L2, but where not specified as having any meaningful effects, so they
! where simply specified as freely correlated to the other variables.
clim_w WITH EF1_W EF2_W ACH1_w ACH2_w;

! controls:
EF1_W ON GENDER FRELUNC;
EF2_W ON GENDER FRELUNC;
ACH1_w ON GENDER FRELUNC;
ACH2_w ON GENDER FRELUNC;
clim_w ON GENDER FRELUNC;
EF1_W WITH ACH1_w;

! Residual variances parameter labels for use with model constraints;

EF1_W (rvr_efw1);
ACH1_W (rvr_acw1);
EF2_W (rvr_efw2);
ACH2_W (rvr_acw2);
! Variances of exogenous variables specified so as to allow them to be taken into account for FIML:
gender;
FRELUNC;

%between% ! This is for the L2 part of the model.
!Time 2 efficacy
EF2_b BY EFF1907@1  (9);
EF2_b BY EFF3107 (10);
EF2_b BY EFF3507 (11) ;
EF2_b BY EFF4407 (12);

!Time 2 Achievement
ACH2_B BY CSTRAW07@1; CSTRAW07@0;

! using parameter constraints to help the model to converge on a proper solution
! forcing uniquenesses to be non-zero
EFF1907 (u5); EFF3107 (u6); EFF3507 (u7); EFF4407 (u8);
MA607 (u9); MA1407 (u10); MA2307 (u11); MA4807 (u12);
CH1307 (u13); CH1507 (u14); CH2707 (u15); CH3707 (u16);
CA2907 (u17); CA3207 (u18); CA4307 (u19);
MA_b (u20); CH_b (u21); CA_b (u22);

!Mastery climate
MA_b BY MA607@1 (1) ;
MA_b BY MA1407 (2) ;
MA_b BY MA2307(3);
MA_b BY MA4807 (4) ;
!Challenge climate
CH_b BY CH1307@1   (5);
```

```
CH_b BY CH1507 (6);
CH_b BY CH2707   (7);
CH_b BY CH3707  (8);
!Caring Climate
CA_b BY  CA2907@1   (13);
CA_b BY CA3207 (14);
CA_b BY CA4307  (15);
!Second order climate
clim_b BY MA_b@1 (16);
clim_b BY CH_b (17);
clim_b BY CA_b (18);

! Regression of achievement on efficacy
ACH2_b ON EF2_b (b_a2e2_b);

! Climate predicts efficacy:
EF2_b ON CLIM_b (b_EFFCL);

! For a test of PARTIAL MEDIATION, add these:
! ACH2_B ON CLIM_b (b_ACHCL);

! Controls
EF2_b ON glev_b;
ACH2_B ON glev_b;
CLIM_b ON glev_b;

! Residual variances parameter labels for model constraints:
EF2_b (rvr_efb2);
ACH2_B (rvr_acb2);
CLIM_b (rvr_CLb2);

MODEL CONSTRAINT:
u5 > 0; u6 > 0; u7 > 0; u8 > 0; u9 > 0; u10 > 0; u11 > 0; u12 > 0; u13 > 0; u14 > 0;
u15 > 0; u16 > 0; u17 > 0; u18 > 0; u19 > 0; u20 = 0; u21 > 0; u22 > 0;

! The NEW function is used to create new parameters outside of the main model for the model constraints
! Variances estimates for efficacy and achievement used to compute effect sizes (taken from CFA model)
!i.e., .031  is the value of the variance of L2/Time 2 self-efficacy in the multilevel CFA model.
NEW (varef2_b); varef2_b = 0.031 ;
NEW (varef2_w); varef2_w = 0.427;
NEW (varac2_b); varac2_b =  0.102;
NEW (varac2_w); varac2_w = 0.906;

! To properly estimate the standardized and effects sizes of the between (L2) effect
! (contextual, standardized, ES) , computed according to the specifications provided in appendix 2
! and Marsh et al. (2009)

! effect of efficacy --> achievement (contextual effect, see appendix 2)
new(CTEF2AC2); !contextual effect (subtraction of the L1 path from the L2 path using parameter labels)
CTEF2AC2 = b_A2e2_b - b_A2e2_w;
new(stef2AC2);  !Properly standardized contextual effect
stef2AC2= CTEF2AC2*(sqrt(rvr_efb2)/sqrt(varac2_w + varac2_b));
new(ebef2AC2);  !Properly estimated contextual effect-size
ebef2AC2=CTEF2AC2*(2*sqrt(rvr_efb2) /sqrt(varac2_w));

! effect of efficacy --> achievement (simple L2, not properly treated as contextual)
```

```
new(st_e2A2);
st_e2A2= b_A2e2_b *(sqrt(rvr_efb2)/sqrt(varac2_w + varac2_b));
new(ea_e2A2);
eb_e2a2=b_A2e2_b *(2*sqrt(rvr_efb2) /sqrt(varac2_w));

! effect of climate --> efficacy (properly estimated as simple L2, not a contextual effect)
new(st_efCL);
st_efCL= b_EFFCL *(sqrt(rvr_CLb2)/sqrt(varef2_w + varef2_b));
new(eb_efCL);
eb_efCL=b_EFFCL *(2*sqrt(rvr_CLb2) /sqrt(varef2_w));

! effect of climate --> achievement (for tests of partial mediation)
!new(st_acCL);
!st_acCL= b_achCL *(sqrt(rvr_CLb2)/sqrt(varac2_w + varac2_b));
!new(ea_acCL);
!eb_acCL=b_achCL *(2*sqrt(rvr_CLb2) /sqrt(varac2_w));

! In order to compute the indirect effect and confidence intervals based on simple L2 measures,
! not properly treated as contextual (not recommended)
NEW (INDL2);
INDL2 = b_EFFCL * b_a2e2_b;

! In order to compute the indirect effect and confidence intervals taking into account that
! the second path (achievement  --> efficacy) is contextual
NEW (INDCONT);
INDCONT =  b_EFFCL * CTEF2AC2;

OUTPUT: ! to request specific outputs, as defined in the Mplus manual.
SAMPSTAT STANDARDIZED RESIDUAL CINTERVAL MODINDICES (3.0) TECH1 TECH2 TECH4
SVALUES;
```

Appendix 2: Statistical Considerations in the Estimation of Level 2 effects, Appropriate Standardization and Effect Sizes

The distinction between L2 contextual and climate constructs made in the introduction is especially important in doubly-latent multilevel models, as the mathematical specification of these models involves an implicit group-mean centering of the variables and thus results in L2 estimates that are independent from their L1 counterparts but that do not partial out the L1 effects from the L2 estimates (Enders & Tofighi, 2007; Marsh et al., 2009, 2012). It is typical to distinguish between group-mean centering and grand-mean centering in multilevel models (Enders & Tofighi, 2007; Lüdtke, et al., 2008, 2010). In the models considered here, grand-mean centering was used for all constructs included only at L1 (i.e., gender, free lunch, and Time 1 math self efficacy and achievement). For L1 variables aggregated to form L2 variables, there is an implicit group-mean centering inherent in the way that doubly latent models handle the decomposition of L1 and L2 effects. Given this implicit group-mean centering, the effects of the L2 variable are removed from the corresponding L1 variable, but the effect of the L2 variable is not controlled for the L1-effect—both effects are estimated as independent from one another. The appropriate L2 contextual effect (i.e. the relation between L2 self-efficacy and achievement) can be obtained by subtracting the L1-effect from the L2-effect. This allows one to ensure that the L2 effect of the contextual variables really adds something to the effects of the main L1 construct being assessed by the items. This is done by calculating an additional parameter representing the difference between the L2 and L1 coefficients, providing a direct estimate of contextual effects (Enders & Tofighi, 2007; Marsh et al., 2009, 2012), with the multivariate delta method (e.g., Raykov & Marcoulides, 2004). This allows for an L2 estimate identical to what would have been obtained under the grand-mean centering procedure, which results in the estimation of partial regression slopes at L2 that are controlled for the influence of the L1 variable. Similarly, the indirect effect of classroom level climate on math achievement as mediated by math self-efficacy needs to be calculated while taking into account the contextual nature of the effect of self-efficacy

on achievement. These effects are illustrated in the input presented in Appendix 1.

For climate ratings, students within the same class are asked to rate common L2 constructs rather than idiosyncratic characteristics. In this case, L2 classroom climate effect does not need to be adjusted for the corresponding L1 effect, since the main construct being assessed is naturally located at L2 and independent from the L1 counterpart. Indeed, the total rating by each student confounds two different components. The first is the shared agreement that represents the climate effect (the L2 construct). The second is the residual L1 variance that represents unique perceptions of each student that are not explained by the shared perceptions of different students. These two components (the residual L1 variance and the L2 shared perspective that represents climate) are automatically estimated as independent from one another in a group-mean centered multilevel model. This is why the appropriate interpretation of the climate effect is the effect of the L2 variable, not the L1 ratings by individual students. Thus, when studying classroom climate effects, the L1 component of climate variables simply needs to be modeled as part of estimating the level of agreement for students within classes and since their meaning is uncertain, they generally should simply be modeled as correlated with the remaining L1 latent variables. The observation of low correlations between L1 climate variables and the other latent variables would be consistent with the rationale that the critical estimate of climate effects lies in the L2 part of the model.

Standardized effects as operationalized in Mplus are currently standardized separately for each level. This is reasonable when the researcher wants to evaluate these coefficients separately at L1 and L2. However, these default standardized coefficients are not particularly useful since, (i) contextual effects need to consider coefficients differences between the two levels, and (ii) for climate effects, all effects are assumed to occur at L2. Following Marsh et al. (2009), all L1 and L2 effects were first standardized in relation to the total (L1 and L2) variance. Then we computed a measure of effect size as proposed by Marsh et al. (2009) that is comparable to Cohen's *d* (Cohen, 1988): ES = $(2 * B * SD_{\text{predictor}})/ SD_{\text{outcome}}$ where B is the unstandardized regression

coefficient, $SD_{predictor}$ is the standard deviation of the predictor, and $SD_{outcome}$ is the L1 standard deviation of the outcome.

In this study, the relation between L2 classroom levels of math achievement and self-efficacy represented a contextual effect and thus needed to be properly estimated as the difference between the L2 and the L1 effect. To illustrate the impact of this procedure on the results, the properly estimated L2 effect of math self-efficacy on achievement is reported at the end of Table 2 (estimate = .587, SE = .234; $p \leq .05$; standardized estimate = .073, SE = .039; $p \leq .05$; effect size = .154, SE = .060; $p \leq .05$) and the properly estimated indirect effect of L2 classroom climate on L2 math achievement as mediated by math self-efficacy is reported in the text (indirect effect = .359; SE = .158; $p \leq .05$; 95% confidence interval = .049 to .668). In contrast, the non-disaggregated (non-contextual) direct (estimate = .739, SE = .232; $p \leq .05$; standardized estimate = .092, SE = .029; $p \leq .05$; effect size = .193, SE = .061; $p \leq .05$) and indirect (indirect effect = .451; SE = .165; $p \leq .05$; 95% confidence interval = .127 to .776) effects were slightly larger. Although this difference did not impact the substantive interpretations in the context of the present study, the proper results are smaller in magnitude than the non-disaggregated L2 effects, reinforcing the implication that care should be taken to estimate these effects properly. The previously described big-fish-little-pond effect (Marsh 2007a) provides an even more vivid example of the importance of proper estimation of L2 effects based on contextual L2 constructs.

Appendix 3: Detailed results from the second-order multilevel CFA model

Table 1.1

Factor Loadings

	T1 Math self-efficacy		T2 Math self-efficacy		Challenge climate		Caring climate		Mastery Climate		Second-Order Climate	
	Est. (S.E)	Std. (S.E.)	Est. (S.E)	Std. (S.E.)	Est. (S.E)	Std. (S.E.)	Est. (S.E)	Std. (S.E.)	Est. (S.E)	Std. (S.E.)	Est. (S.E)	Std. (S.E.)
Level 1												
Indicator 1	1.000 (.000)	.616 (.018)	1.000 (.000)	.674 (.015)	1.000 (.000)	.381 (.028)	1.000 (.000)	.703 (.019)	1.000 (.000)	.427 (.027)	1.000 (.000)	.969 (.027)
Indicator 2	1.103 (.039)	.680 (.017)	1.103 (.039)	.737 (.014)	1.476 (.108)	.534 (.029)	.805 (.048)	.571 (.030)	1.243 (.086)	.526 (.031)	.697 (.071)	.796 (.027)
Indicator 3	1.236 (.040)	.768 (.015)	1.236 (.040)	.825 (.013)	1.909 (.160)	.693 (.018)	1.024 (.041)	.736 (.018)	1.171 (.090)	.507 (.022)	1.238 (.102)	.767 (.023)
Indicator 4	1.173 (.035)	.725 (.015)	1.173 (.035)	.782 (.012)	1.033 (.090)	.375 (.031)	---	---	1.361 (.111)	.580 (.026)	---	---
Level 2												
Indicator 1	---	---	1.000 (.000)	.790 (.075)	1.000 (.000)	.365 (.049)	1.000 (.000)	.917 (.031)	1.000 (.000)	.702 (.086)	1.000 (.000)	1.000 (.000)
Indicator 2	---	---	1.103 (.039)	.904 (.047)	1.476 (.108)	.771 (.052)	.805 (.048)	.702 (.063)	1.243 (.086)	.932 (.041)	.697 (.071)	.910 (.051)
Indicator 3	---	---	1.236 (.040)	.953 (.039)	1.909 (.160)	1.000 (.000)	1.024 (.041)	.906 (.046)	1.171 (.090)	.720 (.078)	1.238 (.102)	.703 (.062)
Indicator 4	---	---	1.173 (.035)	.956 (.046)	1.033 (.090)	.505 (.072)	---	---	1.361 (.111)	.915 (.039)	---	---

Note. All loadings significant ($p \leq .05$); Est. = unstandardized parameter estimate; S.E.: standard error of the estimate; Std. = standardized parameter estimate.

Table 1.2

Covariances (above the Diagonal), Variances (in the diagonal) and Correlations (below the Diagonal)

Level 1	T1 self-efficacy	T2 self-efficacy	Climate	T1 achievement	T2 achievement	Gender	Free lunch
T1 self-efficacy	.383 (.024)*	.187 (.019)*	.048 (.008)*	.171 (.019)*	.160 (.018)*	-.002 (.008)	.000 (.007)
T2 self-efficacy	.461 (.031)*	.427 (.026)*	.121 (.013)*	.193 (.018)*	.208 (.017)*	.004 (.008)	-.008 (.007)
Climate	.194 (.029)*	.467 (.032)*	.157 (.020)*	.008 (.011)	.012 (.010)	.024 (.006)*	.000 (.006)
T1 achievement	.275 (.025)*	.294 (.025)*	.019 (.028)	1.012 (.031)*	.766 (.026)*	.007 (.010)	-.050 (.014)*
T2 achievement	.271 (.026)*	.335 (.021)*	.033 (.027)	.800 (.012)*	.906 (.031)*	.001 (.011)	-.050 (.013)*
Gender	-.005 (.026)	.012 (.024)	.122 (.027)*	.013 (.021)	.001 (.023)	.250 (.000)*	.002 (.005)
Free lunch	.001 (.026)	-.028 (.023)	-.001 (.030)	-.106 (.028)*	-.113 (.027)*	.008 (.021)	.216 (.007)*

Level 2	T2 self-efficacy	Climate	T2 achievement	Grade
T2 self-efficacy	.031 (.008)*	.024 (.006)*	.018 (.008)*	-.018 (.011)
Climate	.688 (.098)*	.039 (.010)*	.016 (.009)	-.026 (.011)*
T2 achievement	.316 (.120)*	.260 (.137)*	.102 (.018)*	.036 (.017)*
Grade	-.205 (.128)	-.268 (.113)*	-.229 (.105)*	.239 (.005)*

Note. Standard errors in parentheses.
* = $p \leq .05$