EDUCATIONAL PSYCHOLOGIST, *47*(2), 106–124, 2012

ISSN: 0046-1520 print / 1532-6985 online
DOI: 10.1080/00461520.2012.670488

Routledge
Taylor & Francis Group

Classroom Climate and Contextual Effects: Conceptual and Methodological Issues in the Evaluation of Group-Level Effects

Herbert W. Marsh
Department of Education
University of Oxford
Centre for Positive Psychology and Education
University of Western Sydney
College of Education
King Saud University

Oliver Lüdtke
Center for Educational Science and Psychology
University of Tübingen
Department of Psychology
Humboldt University

Benjamin Nagengast and Ulrich Trautwein
Center for Educational Science and Psychology
University of Tübingen

Alexandre J. S. Morin
Department of Psychology
University of Sherbrooke
Centre for Positive Psychology and Education
University of Western Sydney

Adel S. Abduljabbar
College of Education
King Saud University

Olaf Köller
Leibniz Institute for Science and Mathematics Education
University of Kiel

Classroom context and climate are inherently classroom-level (L2) constructs, but applied researchers sometimes—inappropriately—represent them by student-level (L1) responses in single-level models rather than more appropriate multilevel models. Here we focus on important conceptual issues (distinctions between climate and contextual variables; use of classroom L2 rather than student-level L1 measures) and more appropriate multilevel models. To illustrate

Correspondence should be addressed to Herbert W. Marsh, Department of Education, University of Oxford, 15 Norham Gardens, Oxford OX2 6PY, United Kingdom. E-mail: herb.marsh@education.ox.ac.uk

these issues, we consider the effects of two L2 classroom climate variables and one L2 classroom contextual variable on two L1 student-level outcomes for 2261 students in 128 classes. Through this example, we illustrate how to apply evolving doubly latent multilevel models to (a) evaluate the factor structure of L1 and L2 constructs based on multiple indicators of classroom climate and context measures, (b) control measurement error at L1 and L2, (c) control sampling error in the aggregation of L1 responses to form L2 constructs (the average of student-level responses to form classroom-level constructs), and (d) provide guidelines for appropriate analysis of classroom climate as an L2 construct.
[Supplementary materials are available for this article. Go to the publisher's online edition of *Educational Psychologist* for the following free supplemental resources: Substantive basis of the present investigation and more detailed description of the methodology.]

In educational research, contextual and 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 are based on the aggregation of L1 student-level variables (e.g., within-class averages of individual student achievement or socioeconomic status (SES) used to form classroom contextual variables; average within-class student ratings of classroom organization or evaluations of teacher enthusiasm used to form classroom climate variables). This general strategy is at the heart of educational research (e.g., school/teacher effectiveness studies, value-added models, classroom/school climate and contextual studies). However, the conceptual framework presented here is also widely applicable to organizational, family, sociology, and medical research (e.g., Bliese, 2000; Bliese, Chan, & Ployhart, 2007; Croon & van Veldhoven, 2007; Iverson, 1991; Kozlowski & Klein, 2000; LaHuis & Ferguson, 2009; Shin & Raudenbush, 2010). In fact, these issues are central to any area of research in which individuals interact with other individuals in a group setting, leading Iverson (1991) to conclude, "This range of areas illustrates how broadly contextual analysis has been used in the study of human behavior" (p. 11)—how group-level characteristics influence individual outcomes.

Particularly in educational climate and contextual studies, however, there is widespread confusion about the appropriate nature of data, design, statistical models, and interpretations. Fundamental design and analytic problems in many published studies seriously undermine substantive interpretations of the results (also see subsequent guidelines presented at the end of this article). Following from Lüdtke, Robitzsch, Trautwein, and Kunter (2009), we argue that there are three critical methodological issues that contextual and climate studies must address: (a) the appropriate unit of analysis (student, classroom, or both)—the primary interpretations must be based on appropriate aggregations of L1 responses to form L2 constructs, not on single-level models of L1 individual student ratings; (b) statistical models that control for measurement error at L1 and L2, and sampling error in the aggregation from L1 to L2; and (c) multilevel models that appropriately operationalize and distinguish between climate and contextual effects. Historically, many applied researchers have not had the statistical tools or, apparently, an appropriate understanding of multilevel data to meet these challenges. In particular, basic unit-of-analysis problems are widespread in published research; treating individual (L1) students' ratings of classroom climate as if they were the L2 classroom constructs. However, even among multilevel studies, there are few studies of contextual or climate effects in the applied educational research literature that meet these challenges.

A PRIMER OF KEY TERMS AND CONCEPTUAL ISSUES

The purpose of this section is to introduce terminology and issues that are expanded upon later in the article, and to illustrate these with examples relevant to educational psychology. In classroom research it is important to distinguish between L1 constructs based on individual student-level (L1) constructs and classroom-level (L2) constructs. L1 constructs are obviously based on responses by individual students. However, L2 classroom-level constructs can be based on true L2 measures (e.g., the number of students in a class, the gender of the teacher) or aggregates of responses by students within the class (e.g., class-average achievement levels—a contextual variable; class-average ratings of teacher friendliness—a climate variable). The evaluation of classroom climate or contextual variables should always be based on L2 classroom-level constructs, not L1 student-level responses. Although similar in some respects, the key distinction between climate and context variables is the referent in the L1 measure. For classroom climate constructs (e.g., teacher friendliness or classroom organization), the referent is the classroom (or teacher) in that each student in the class rates some aspect of the class or teacher, not some individual characteristic of the student making the rating. In this sense the ratings of individual students are interchangeable in relation to scores reflecting the L2 classroom climate variable in that every student is instructed to rate the same construct. Hence, ideally there should be very good agreement among students within the same

class. For classroom context constructs (e.g., class-average achievement or gender ratio), the referent is the individual student and the L2 construct is an aggregation of these different student characteristics. Here the different students are not interchangeable in the sense that different students within the same group have different true scores. We elaborate this distinction and its implications later in the article (also see subsequent guidelines presented at the end of this article).

Particularly in contextual studies, the same L1 variable can be used to construct both L1 and L2 constructs, but they can have fundamentally different interpretations. For example, a well-documented finding in self-concept research is the big-fish-little-pond effect (BFLPE; e.g., Marsh, Seaton, et al., 2008) whereby individual student (L1) achievement has a *positive* effect on academic self-concept (the brighter I am, the better my academic self-concept), but classroom-average (L2) achievement has a *negative* effect on academic self-concept (the brighter my classmates are, the lower my academic self-concept) after controlling for individual achievement. This is an example of a contextual variable in which it is critical to simultaneously consider both L1 and L2 constructs, as the appropriate interpretation of one depends on the other. This is most appropriately accomplished with multilevel analyses that incorporate both student- and classroom-level variables into one statistical model—rather than single-level models that consider only the student level. Classroom climate and contextual constructs are inherently classroom-level (L2) constructs and should be represented as such in statistical models.

In educational psychology there have been two dominant developments in statistical analyses. Confirmatory factor analysis, structural equation modeling, and related statistical approaches are primarily concerned with latent variables based on multiple indicators; these methods address the factor structure of how indicators are related to the latent variables (factors) they are intended to represent, issues of measurement error, and relations among latent variables that control for measurement error. Particular foci of these analyses are tests of the a priori factor structure to fit the data (based on factor loadings and other parameter estimates as well as goodness of fit indices), correction of relations among factors for measurement error, and support for the construct validity of the factors based on the factor structure and relations among the factors. However, factor analytic approaches have traditionally been single-level analyses based on responses by individuals—not multilevel studies that simultaneously consider L1 individual student and L2 classroom constructs.

Multilevel modeling and related techniques are primarily concerned with appropriately representing the multilevel nature of educational data in which students are nested under classrooms, classrooms are nested under schools, schools are nested under school districts, and so on. There are statistical reasons why it is important to use multilevel modeling when the data have a multilevel structure (e.g., nonindependence of responses by students within the same classroom that could result in inflated Type 1 error rates). However, our focus here is on conceptual issues and the need to simultaneously consider variables at the (L1) student level and (L2) classroom level in the same statistical model as illustrated in the earlier example of the BFLPE. Particular foci of these analyses are the appropriate decomposition of variables into variance components at different levels (e.g., how much variance in student-level responses is explained by the student's classroom or school) and relations between variables at different levels of analysis (e.g., how individual and classroom-average achievement are related to individual academic self-concept). Nevertheless, multilevel approaches have traditionally been based upon scale scores (single indicators) of each construct (also known as "manifest" variables) that ignore the multiple indicators upon which they are based, the factor structure relating indicators to factors, and measurement error.

Fortunately, recent advances in statistical modeling provide much stronger tools for educational researchers to evaluate contextual and climate effects. Hence, the overarching purposes of this article are to discuss basic problems in many existing climate and contextual studies and to present a more appropriate conceptual and analytic framework for such studies. Our approach integrates confirmatory factor analysis (CFA), structural equation modeling (SEM), and multilevel modelling (MLM) into a unified statistical framework. In our work we used this integration to develop a doubly latent model of contextual effects (Lüdtke, Marsh, Robitzsch, & Trautwein, 2011; Lüdtke et al., 2008; Marsh et al., 2009; Nagengast & Marsh, 2011). As used here, models are latent when they use multiple indicators that are designed to reflect a hypothetical unobserved construct that is assumed to be responsible for the covariance among the indicators, thus allowing for the control of measurement error, rather than a single measure that is taken to be the construct that has no error. Typically the multiple indicators are multiple items sampled from a potentially large number of possible items. However, the multiple indicators might also refer to multiple persons sampled from a potentially large number of persons. Combining these two usages of the term *latent*, our doubly latent model is

- latent in relation to *measurement error*, as in the traditional factor analysis approach. It incorporates multiple indicators (e.g., items) of each construct to form latent factors that are corrected for measurement error at both the L1 student level and the L2 classroom level. Estimates of measurement error are smaller when correlations among the multiple indicators are higher (i.e., agreement among the items is better) and when there are more (equally appropriate) items—a traditional approach to reliability analysis. A minimal requirement for support of the latent factor is that agreement among the different items designed to measure the same factor is higher than agreement among items designed to measure different factors.

- latent in relation to *sampling error*, as in the traditional multilevel approach. The doubly latent model incorporates 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. Estimates of sampling error are smaller when agreement among the different students in the same class is better and when the number of students is larger (as in traditional measures of intraclass correlations [ICCs]). In this sense, the construct is latent because it uses responses from multiple students to control for sampling error. Thus, for example, the mean achievement scores from a sample of students from the same school is a sample mean with a degree of uncertainty due at least in part to sampling error, not a population mean. A minimal requirement for support of the latent classroom construct is that agreement among the students within the same class is higher than agreement among students from different classes.

In this respect, a similar logic is used to identify latent constructs on the basis of multiple indicators—the aggregation of items to form latent factors and the aggregation of scores by different students within the same class to form latent classroom-level constructs. In each case it would be possible to use simple unweighted means (the average item response, a scale score; the average student response, a classroom average). However the doubly latent multilevel model provides information about the structure of responses to evaluate the latent constructs and correction of measurement and sampling error. If the latent constructs are well defined and there is little sampling or measurement error, then the use of simple mean scores might be justified, but these typically unrealistic assumptions cannot even be appropriately evaluated without applying the logic of the doubly latent model.

The critical feature for both classroom contextual and climate variables is that they should always be conceptualized as classroom-level constructs that are most appropriately evaluated from a multilevel perspective. Although the doubly latent model has been applied to contextual studies, here we extend it to include both climate and contextual variables. Although the logic of the doubly latent model applies to both contextual and climate variables, there are also important distinctions between the two.

Distinctions Between Climate and Contextual Constructs

It is critical that educational researchers understand that contextual and climate effects should be based on group-level measures, either true L2 constructs or aggregates of L1 measures to form L2 constructs that are the focus of this article. However, it is also important to distinguish between contextual and climate variables. This conceptual distinction is critical in terms of how contextual and climate variables are designed and operationalized, how they are tested in doubly latent models emphasized here, and how they are interpreted. Nevertheless, there is much confusion about these issues in the educational psychology research literature.

Definition, design, and operationalization of contextual and climate variables. Contextual effects are the effects of an L2 classroom-level variable above the effect of the corresponding L1 variable upon which it is based (e.g., the effect of L2 class-average achievement after controlling for the effect of L1 individual student achievement). Contextual constructs are classroom L2 aggregates of individual student L1 characteristics that are specific to each person in the class (e.g., class-average achievement, class-average SES, student gender). Here the referent is the individual student rather than the class, and class averages are used to describe classroom composition (i.e. context). That is, contextual variables are based on aggregations of an L1 construct that is specific to the person, and not interchangeable with other students. For example, each student within a class is either male or female (an L1 characteristic of the individual student that is clearly not interchangeable across different students in the same class), and a measure of the gender composition of each class might consist of the percentage of males or females in the class (an L2 classroom contextual variable). Here the referent of the contextual variable is the gender of each individual student—not the class as a whole—and the gender of each student is not interchangeable with the gender of other students in the same class. Thus, classroom gender composition is a contextual variable rather than a climate variable.

Climate constructs are classroom (L2) aggregations of ratings by students when each student is asked to rate some characteristic of the group or classroom that is common to all students (e.g., classroom orientation, teacher enthusiasm). Here the referent is the classroom, not the individual student. That is, the referent is the same for each student within the same class; each student directly rates the same L2 construct and not some individual L1 characteristic that is specific to the person making the rating. In classroom climate ratings the class referent is typically made explicit. For example, scales and items from the widely used Pattern of Adaptive Learning Scales instrument (Midgley, 2002) include 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"). The main purpose of the L1 climate ratings is to assess an L2 construct, not individual student characteristics. Thus the referent for classroom climate variables is the whole class.

For aggregated climate variables, all students within the same class rate the same L2 construct, and the classroom climate is based on the shared perceptions of students within

the class. In this respect the group members are theoretically interchangeable (in that all students within the same class are rating the same classroom climate) and students are directly rating the L2 construct. From this perspective, classroom climate is based on the shared perceptions among different students within the same class, whereas differences among students within the same class (residual L1 differences after controlling for shared agreement) are a source of unreliability in the L2 climate construct. This is not to say that there are no systematic individual differences among the ratings by L1 students within each class, but merely that these individual differences do not reflect the L2 classroom climate of interest (i.e., the shared agreement among students from the same class). This point was made in the classic 1976 article by Cronbach, who noted that studying individual differences in the perceptions of different students within the same classroom might be interesting but does not reflect classroom climate. From this conceptual perspective, it follows that if there is no agreement among students within the same class in relation to the classroom climate variable, then the aggregated measure of climate is completely unreliable and probably should not be considered further.

The L1 ratings of climate are important in terms of estimating agreement among students within the class and forming the L2 aggregates, and they may be related to other L1 constructs. However, the residual L1 climate ratings have no substantive meaning in relation to the interpretation of L2 climate effects. Rather, they represent unique perceptions of each student that are not explained by the shared perceptions of different students. Indeed, ideally there should be little or no systematic residual variance at the L1 level beyond what can be explained by the L2 latent climate factor. The theoretical and statistical rationale for this claim dates back to the seminal Cronbach (1976) paper on multilevel models but is still widely misunderstood in classroom climate research.

THE METHODOLOGICAL FOCUS: DOUBLY LATENT MULTILEVEL MODELS OF CLIMATE AND CONTEXTUAL EFFECTS

As emphasized earlier, educational research is inherently multilevel; students are nested within classrooms, classrooms are nested within schools, schools are nested within school systems, and so forth. Characteristics of the school, classroom, and teacher are often assessed via student reports and related to outcome variables such as student achievement or student motivation (Lüdtke et al., 2009). When classroom climate and contextual effects are based on the perceptions of individual students, they should be studied as classroom-level (L2) constructs from a multilevel perspective. Critical issues in studies of contextual and climate effects include the use of the appropriate unit or level of analysis, multilevel statistical models that control for measurement and sampling error, and operationalization of the distinction between contextual and climate effects.

Level of Analysis

The appropriate level (or unit) of analysis and interpretation depends, at least in part, on the research questions and the nature of the data collected. If researchers are interested in the effects of L2 classroom (teacher or school) variables on other L2 variables, then the appropriate unit of analysis for the interpretation of these effects is the classroom (teacher or school). However, complications about the appropriate unit of analysis arise when the data contain a mixture of L1 student-level responses and L2 constructs—particularly when the outcomes include both L1 student-level variables and L2 classroom-level constructs, which are based on aggregates of L1 responses by individual students. Nevertheless, even in such situations, the primary focus should still be on the effects of L2 constructs based on the aggregation of responses by L1 students within the same classroom—and their effects on L1 student-level outcomes or, perhaps, on other L2 classroom-level outcomes.

Methodologically, this issue of the relevant unit of analysis has been well understood for more than a quarter of a century. In his seminal paper on multilevel issues in educational research in relation to classroom climate, Cronbach (1976) argued,

> 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. (p. 18)

Yet, in their review of classroom goal structures, Miller (2006; Miller & Murdock, 2007) found that 16 of 31 studies did not even consider any class-level analyses. Like Cronbach and many others (Lüdtke et al., 2008; Lüdtke et al., 2009; Marsh et al., 2009; Papaioannou, Marsh, & Theodorakis, 2004), Miller warned of the inappropriateness of this analytic approach to evaluating classroom climate. Although many recent studies have emphasized the need to measure classroom climate at the classroom level (e.g., Mainhard, Brekelmans, & Wubbels, 2011), many publications in leading educational journals continue to treat the raw L1 measures (rather than L2 aggregates of L1 measures) as if they are climate measures (e.g., Ciani, Middleton, Summers, & Sheldon, 2010; Fast et al., 2010; Wang & Holcombe, 2010). Indeed, Ciani et al. (2010) went so far as to argue—inappropriately—that

> when attempting to predict between-class variance in a motivational outcome variable with classroom goal structure, researchers often use the classroom aggregate of student

> perceptions of classroom goal structure as the level 2 predictor variable. While not an entirely inappropriate technique, we believe that student perceptions are better suited as level 1 predictors. (p. 97)

That is, as noted earlier and emphasized by Cronbach (1976) and many others subsequently (e.g., Lüdtke et al., 2008; Lüdtke et al., 2009; Marsh et al., 2009; Miller, 2006; Papaioannou et al., 2004) it is the L2 aggregation of L1 student responses that represents classroom climate, not the ratings by individual students. Hence, despite the clear resolution of this methodological issue for more than a quarter of a century, it is still an area of ongoing confusion in the educational literature.

Psychometric Properties of Aggregated Student Ratings

In social science research there is a long history of assessing the psychometric properties of L1 measures based on multiple indicators. This focus on multiple indicators is particularly evident in CFAs and SEMs that simultaneously evaluate the goodness of fit based on a priori factor structures, control for unreliability, and evaluate construct validity based on substantive interpretations of parameter estimates. In classroom research, there is a conceptually similar issue in which L1 responses by different students in the same class are the multiple indicators of the L2 classroom climate construct. Although there is a long tradition in educational research of evaluating the generalizability of class means (see Brennan, 2001; Hoyt, 2000; Kane, Gillmore, & Crooks, 1976) based on achievement test scores, this is not typically pursued in classroom climate and contextual research. However, Lüdtke et al. (2009) argued that it is important to evaluate the psychometric properties of L2 aggregates of L1 student ratings and to determine whether it even makes sense to form aggregate variables in the first place.

When L1 measures are aggregated to form L2 measures, there is a second source of uncertainty—sampling error, in addition to measurement error associated with the multiple items. We use the term *sampling error* to refer to the uncertainty introduced by sampling L1 students when their ratings are used to form the L2 construct. As emphasized by Marsh (2009; also see Lüdtke et al., 2008; Lüdtke et al., 2011; Shin & Raudenbush, 2010), the observed group mean can only be used to infer a corresponding population mean with some degree of uncertainty. This uncertainty may be because only a sample of the individuals from a group was measured or because the individuals making up the group are assumed to be a sample of individuals from a broader population. As emphasized by Shin and Raudenbush (2010),

> The sample mean of the covariate is generally used to represent the organization [the L2 unit] mean on the covariate. Unfortunately, the sample mean is an unreliable estimate of the organization mean, and this unreliability will generally lead to bias not only in estimating the contextual effect but also in estimating the association between other organizational covariates and the outcome controlling for the contextual effect. (p. 27)

Measures of reliability used to assess measurement error at L1 or L2 (e.g., coefficient alpha) depend upon the average agreement among the items that are aggregated to form a score and the number of items. Agreement among items can be indexed as the average correlation among items. However, even when the average correlation among items designed to measure the same construct is only modest, the overall score can be highly reliable if there is a sufficiently large number of items.

The same logic can be applied to the assessment of sampling error when L1 ratings by individual students are aggregated to form L2 aggregates of classroom climate. Sampling error for L2 classroom variables formed from L1 responses is a function of the average agreement among students in the same class and the number of sampled students in each class. Even when agreement among students is only modest, the L2 aggregate can be highly reliable if there are enough students rating each class. 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 ICC1 and is sometimes referred to as the single-rater reliability. 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 and ICC2 as the reliability of the group average (analogous to the reliability of a factor based on multiple items; see Lüdtke et al., 2008; Lüdtke et al., 2011; Marsh et al., 2009; also see the online supplementary materials, Appendix 2). Thus, a modest ICC1 (like a modest average correlation among items) can result in a high ICC2 (i.e., a highly reliable class-average mean based on the aggregate of individual student ratings) if there are enough sampled students in the class (in the same way that a modest average correlation among items can result in a highly reliable factor if there are enough items).

In applied educational research, there is widespread application of (single-level) SEMs with multiple indicators of individual-level constructs (e.g., Jöreskog & Sörbom, 1988; Kaplan, 2000; Marsh, 2007a; Marsh, Byrne, & Yeung, 1999) and of MLMs based on (single) manifest indicators of each construct (e.g., Goldstein, 2003; Raudenbush & Bryk, 2002). Nevertheless, progress has been slow in integrating these two dominant analytical approaches into a single framework in a way that they can be easily implemented in applied research—the primary focus of the present article. Early developments (e.g., Goldstein & McDonald, 1988; McDonald, 1993, 1994; also see Goldstein, 2003) laid the foundation for important advances, but these were not easily implemented with existing software (e.g., McDonald, 1994). B. O. Muthén (1989, 1994) demonstrated multilevel SEM

applications using his partial maximum likelihood estimator and subsequently implemented a full information likelihood estimation procedure. Rabe-Hesketh, Skrondal, and Pickles (2004) also demonstrated how this can be accomplished in their generalized linear latent and mixed models framework and argued that

> a synthesis of both methods, namely multilevel structural equation models, is required when the units of observation form a hierarchy of nested clusters and some variables of interest cannot be measured directly but are measured by a set of items or fallible instruments. (p. 168)

Marsh et al. (2009; Lüdtke et al., 2011) introduced a doubly latent multilevel structural equation model (ML-SEM) for contextual effects that controls measurement error at the student level and the group level as well as sampling error due to the aggregation of L1 variables to form L2 variables (also see Goldstein & McDonald, 1988; McDonald, 1993, 1994; Metha & Neale, 2005; Rabe-Hesketh et al., 2004). The model entails the integration of CFA models to evaluate the factor structure of constructs based on multiple indicators of constructs at L1 and L2; this provides a control of measurement error at both levels of analysis. In addition, all group-level (L2) variables that are based on aggregation of individual students (L1) variables are formed by latent aggregation, taking account of sampling error in going from student-level variables to class-level aggregates. However, apart from didactic examples (e.g., Lüdtke et al., 2011; Lüdtke et al., 2008; Marsh et al., 2009), these recent advances have not yet been fully implemented into published contextual studies (but see Nagengast & Marsh, 2011) and apparently have not been applied to classroom climate studies. Hence, the overreaching purpose of the present investigation is to demonstrate the application of these doubly latent ML-SEMs to contextual studies and their extension to classroom climate studies. Although we demonstrate these statistical models with an exemplary data set, our focus is on the conceptual and methodological issues underpinning these statistical models that have wide applicability to educational psychology and the social sciences more generally.

THE PRESENT DEMONSTRATION

Our article is designed to illustrate important conceptual issues and demonstrate key features of evolving statistical procedures used to evaluate classroom climate and contextual effects. We demonstrate the application of the doubly latent ML-SEMs proposed by Marsh et al. (2009; Lüdtke et al., 2011; Lüdtke et al., 2008) to the study of classroom contexts and its extension to classroom climates. This approach integrates traditional CFA/SEM approaches used to evaluate support for a priori factor structures and to control for measurement error and MLM approaches used to control for sampling error and to unconfound the effects at the level of individual students and their classrooms. Key constructs are inferred on the basis of responses by individual students (L1) to multiple indicators, providing a correction for measurement error at L1 (as in traditional CFA models). Climate variables at the L2 classroom level are based on a latent aggregation of these student-level responses to form multiple indicators of classroom constructs, thus controlling for sampling error (as in traditional multilevel models). Based on the multiple indicators at L2, latent classroom variables are doubly latent, controlling for measurement error at both the individual student and classroom levels, and controlling for sampling error in the aggregation of individual student responses to form classroom-level constructs.

Data used in this demonstration are from the extended German sample from the Third International Mathematics and Science Study (Baumert et al., 1997) and are used here in a reanalysis and extension of the Lüdtke, Köller, Marsh, and Trautwein (2005) study where the sample (2,261 students in 128 classes collected in Grades 7 and 8) and procedures are described in detail. Math achievement (MAch) in Grades 7 (T1) and 8 (T2), and IQ (T1 only) were based on standardized achievement tests. SES was based on mother's and father's education. Math self-concept (MSC) was assessed with a four-item scale that has been validated in many large-scale German studies (e.g., Möller & Köller, 2001; Trautwein, Lüdtke, Marsh, Köller, & Baumert, 2006). Two climate variables were based on individual (L1) student perceptions aggregated to the classroom level (L2), each based on responses to four-item scales: (a) social comparison focus (SCF) climate in which feedback by the teacher emphasizes comparisons with other students rather than progress by the individual student, and (b) classroom chaos climate (CCC), which assesses time wasting and class disruptions. The one contextual variable was based on the class-average (L2) of individual (L1) student scores on the standardized test of mathematics. There are complete data for all measures of IQ and achievement and nearly complete data for all Time 1 psychological constructs. However, because there were considerable missing data for the T2 psychological constructs (21–23%), we implemented the full-information-maximum-likelihood approach to missing data in Mplus (L. K. Muthén & Muthén, 2006–2010; also see Little & Rubin, 1987; Schafer, 1997). All statistical analyses were done with Mplus (version 6; L. K. Muthén & Muthén, 2006–2010) using the doubly latent ML-SEM described in greater detail by Marsh, Lüdtke, and colleagues (Lüdtke et al., 2011; Lüdtke et al., 2008; Marsh et al. 2009; also see the online supplementary materials, Appendix 2 for additional technical details and Appendix 4 for Mplus syntax).

Substantive Basis of the Present Investigation

BFLPE: A contextual effect. The BFLPE is a widely studied contextual effect in educational research (Marsh,

2007b; Marsh & Craven, 2006; Marsh, Seaton, et al., 2008). The BFLPE is a classic contextual effect in which the effect of individual student achievement (L1-ACH) on academic self-concept (L1-ASC) is positive but the corresponding effect of group-average (school or classroom) achievement (L2-ACH) is negative after controlling for individual achievement. Students evaluate their accomplishments in relation to those of their classmates and use this comparative self-evaluation to determine their ASCs. The BFLPE is a robust, long-lasting contextual effect that generalizes across diverse research settings, levels of education, and cultures from all over the world (Marsh, Seaton, et al., 2008). From a policy perspective, the BFLPE provides an alternative, contradictory perspective to educational policy on the placement of students in special education settings, one that is being enacted in many countries throughout the world. The juxtaposition of the positive effects of individual achievement and the negative effects of class-average achievement is inherently a multilevel issue that cannot be represented adequately at either the individual or the classroom level. Hence, it is important to analyze data with appropriate multilevel statistical procedures. Demonstrating the synergy between applied research and methodology, methodological limitations in existing BFLPE research led to the development and extension of the doubly-latent multilevel model demonstrated here (Marsh et al., 2009; Marsh, Seaton, et al., 2008).

Social comparison focus (SCF) and classroom chaos: Climate effects. Particularly in German research there is a substantial classroom climate literature that focuses on the nature of student feedback given by teachers to individual students (e.g., Heckhausen & Heckhausen, 2008; Lüdtke et al., 2005; Rheinberg & Krug, 1999; also see related work by Ames, 1992; Covington, 2001; Marshall & Weinstein, 1984). Rheinberg (1980) distinguished teachers who prefer a social-comparison standard from teachers who prefer an individual standard. Teachers using an individualized standard provided temporal feedback to students and emphasize improvement, whereas teachers with a social-comparison frame of reference assess their students' accomplishments on the basis of comparisons with others. The central characteristic of an individual frame of reference when assessing students' accomplishments is the emphasis on the intraindividual improvement of individual students, thereby enhancing students' self-concept, motivation, and achievement. Subsequently, several studies—including correlational, experimental, and longitudinal designs—have indeed found positive effects of an individualized frame of reference on students' attitudes, attributions, and motivation (e.g., Heckhausen & Heckhausen, 2008; Mischo & Rheinberg, 1995; Rheinberg, 1980; Rheinberg & Krug, 1999). More specifically, Lüdtke et al. (2005) demonstrated that an individualized teacher frame of reference enhanced ASC.

Classroom management is a basic function performed by teachers that influences instructional activities and student outcomes (Brophy, 1988; Evertson, Emmer, Sanford, & Clements, 1983; Lewis, 2001; Matheny & Edwards, 1974). Thus, Brophy emphasized the importance of the

> teacher's ability to maximize the time that students spend actively engaged in worthwhile academic activities, to minimize the time that they spend waiting for activities to get started, making transitions between activities, sitting with nothing to do, or engaging in misconduct. (Brophy, 1988, p. 3)

Rutter, Maughan, Mortimore, Ouston, and Smith (1979) emphasized that more organized teachers, who did not waste time in transitory periods (e.g., handing out papers), had better behaved students. Helmke, Schneider, and Weinert (1986) reported that efficient use of time was positively correlated with student achievement. In summary, chaotic classroom climates are likely to be antithetical to student behavior and outcomes.

Illustration of a Doubly Latent Model of Climate and Contextual Variables

It is useful to use Figure 1 to illustrate how this doubly latent ML-SEM incorporates many of the important features of single-level SEM and multilevel models. For purposes of this model, Time 1 constructs are pretest variables measured when students were in Grade 7, whereas the contextual, climate, and outcome variables of interest are measured at T2 when students were in Grade 8. Each of the indicators of these T2 constructs is depicted as boxes, representing a single measured indictor. Each T2 indicator is decomposed into components at the L1 student level and the L2 class level: the small (L1) ovals below the boxes and the small (L2) ovals above the boxes. This is the traditional MLM approach with its emphasis on the decomposition of effects at L1 and L2. For the two climate variables and the self-concept outcome, there are multiple indicators associated with each construct that are posited to represent a single latent construct. This is the traditional CFA approach with its emphasis on the tests of goodness of fit in relation to an a priori factor structure and correction for measurement error based on agreement among multiple indicators (items). However, this factor structure is replicated at L1 and L2 and this a priori structure can be evaluated (in relation to parameter estimates and goodness of fit).

Tests of contextual and climate variables in the multilevel statistical model and their interpretation. In the doubly latent ML-SEM statistical model presented here, both contextual and climate variables are group-level L2 constructs based on aggregations of individual student-level L1 scores. However, as emphasized earlier, there are important differences in the way climate and contextual variables are

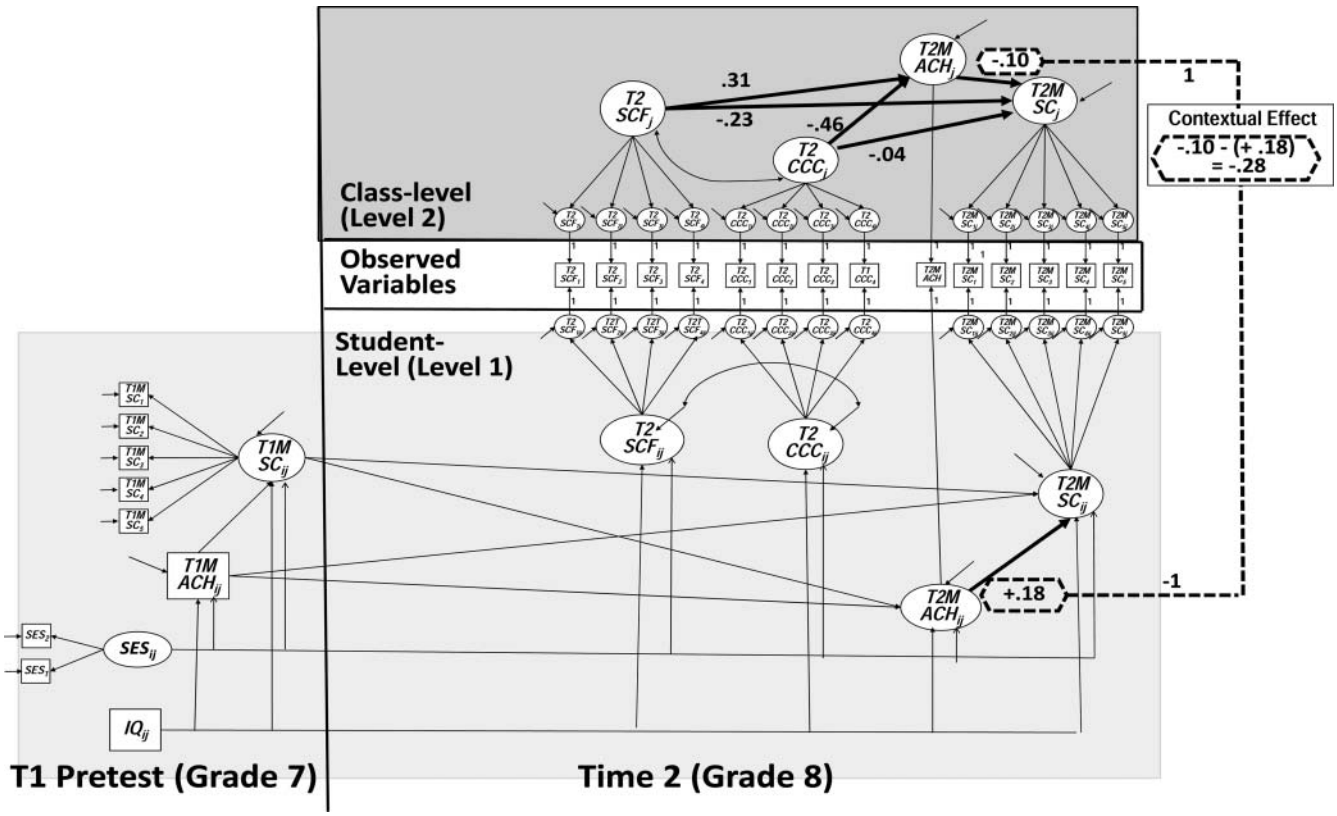

FIGURE 1 Variables on the left side of the diagram are Time 1 (T1) pretest variables collected in Year 7. *Note.* Latent constructs are represented as circles and indicators of these latent variables are represented as squares. For this demonstration, all pretest variables are measured at the individual L1 student level. Math achievement (T1MAch) and IQ are each inferred on the basis of a single total score (based on standardized tests). SES is based on two indicators (mother's and father's education). Math Self-concept (T1MSC) is measured by responses to five items. The subscripts *ij* indicate that these L1 variables take on different values for each student *i* in each classroom *j*. Variables on the right side of the diagram are T2 variables collected in Year 8. The observed variables are more complicated in that they reflect the effects of both individual students (L1) and classrooms (L2). That is, the individual boxes representing the observed variables at T2 are associated with both individual student constructs (L1, those in lighter gray) and classroom (L2, those in darker gray) constructs. Of particular relevance are the two class-average climate variables measured (social comparison focus [SCF] and chaos class climate [CCC]) and the contextual variable (class-average math achievement, T2MAch). Also see the online supplementary materials, Appendix 3, for other parameter estimates and Appendix 4 for Mplus syntax.

measured. In Figure 1 we illustrate how these differences are operationalized in these doubly latent ML-SEMs.

For climate variables, the effect of climate is the effect of the L2 climate variable on the L2 outcomes. In Figure 1 these are the bolded paths from L2 social comparison focus at Time 1 (T2SCF) and chaos class climate at Time 2 (T2CCC) variables to math achievement (T2MAch) and math self-concept (T2MSC). Climate variables are based on the shared agreement among students within the same class, not the L1 rating. Indeed, the total (original, raw, uncontrolled) L1 ratings by each student confound 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 perspectives that represent climate) are automatically disentangled in the 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. Indeed, as the L1 construct in this model is the difference between the rating of an individual student and the class-aggregate rating of all students in the same class, there should be little systematic variation in ratings by individual students if there is very good agreement among students within the same class regarding classroom climate. However, if there is no agreement among students within the same class, then there is no defensible measure of classroom climate (in the same way that if there is no correlation among multiple items designed to measure a construct, then there is no defensible measure of a construct).

For contextual effects, the critical parameter is the L2 effect after controlling for the corresponding L1 effect; the effects of class-average achievement after controlling for the effects of individual student achievement. In Figure 1, this is operationalized as the difference between the L2 and the L1 effects (shown in Figure 1 as the hexagon with dashed lines leading to the L1 and L2 measures of math achieve-

ment). Its standard error (to test its statistical significance) and transformations into standardized values and an effect size (ES) metric (effects standardized in relation to standard deviations) can then be obtained (Marsh et al., 2009; also see the online supplementary materials, Appendix 2; also see the online supplementary materials, Appendix 4 for Mplus syntax), as discussed in the next section.

APPLICATION AND INTERPRETATION OF THE DOUBLY LATENT ML-SEM

Preliminary Results: Reliability of Climate and Contextual Variables

Prior to evaluating contextual and climate effects, it is important to evaluate ICC1s (the amount of variability located at the higher level, an index of the average agreement between pairs of students within the same class) and ICC2s (the reliability of the class-average construct). The ICC1 is an index of the average agreement between any two students in the same class. It is also a variance component—the portion of variance that can be explained by differences between classes. That is, large systematic differences between classes imply that there is more agreement among students within the same class than between students from different classes. If this value is not significantly different from zero or is close to zero, it means that there is little or no systematic differentiation between classes in terms of the climate or contextual variable. In multilevel studies, ICC1s for climate variables are often less than .10 and rarely greater than .30 (e.g., Bliese, 2000; Marsh, Martin, & Cheng, 2008) but can be substantially larger for contextual variables such as class-average achievement, particularly if students are assigned to classes on the basis of their ability levels (e.g., Marsh et al., 2009; Trautwein, Lüdtke, Köller & Baumert, 2006; Trautwein, Lüdtke, Marsh, et al., 2006).

ICC2s depend on ICC1 and the number of students in each L2 class—the higher ICC1 and the larger the number of students, the more reliable the climate or contextual variable is in relation to sampling error. For math achievement at T2 (T2MAch), ICC1 was .558, whereas ICC2 was .957. For classroom chaos climate, ICC1 was .19, whereas ICC2 was .81. Similarly, for social comparison climate, ICC1 was .19, whereas ICC2 was .80. The large ICC1 for achievement reflects substantial differences between classrooms in terms of class-average achievement that are typical in school systems (like Germany) where classes and schools are segregated in relation to achievement. The class-to-class variation in the climate variables is substantially less than for the achievement scores.

In summary, there were statistically significant and meaningful large differences between classes in relation to both the contextual and climate variables, and a sufficiently large number of students in each class to ensure that these differences between classes on the L2 climate and contextual variables are at least moderately reliable. However, the evaluation of ICC1 and ICC2 should be a standard starting point for all multilevel contextual or climate studies (see guidelines presented at the end of this article). If the ICC1 for a climate or contextual variable is statistically nonsignificant or close to zero, the appropriate interpretation is that there is little or no class-to-class variation in the scores used to represent the climate or contextual variables. Although this could mean that either there really were no differences or the measures being used were not appropriate, it makes little sense in either case to pursue further analyses based on these L2 constructs.

Measurement Model and Relations Among Constructs

The design, rationale, and details of statistical analyses for this demonstration are summarized in Figure 1 (also see the online supplementary materials, Appendixes 1–4 for additional results). In the initial set of multilevel CFA models, we tested whether the factor loadings were the same (i.e., invariance of factor loadings) across the multiple indicators at the student (L1) and classroom (L2) levels, and over the multiple indicators measured at time (T1 and T2). These tests were necessary to verify whether the structure of the measured constructs was the same across levels or measurement points. Particularly for fit indices that take into account model parsimony, there was no decrement in fit due to the imposition of any of these invariance constraints (see further discussion in the online supplementary materials, Appendix 3). Thus, in the final measurement model, factor loadings for math self-concept are constrained to be invariant across T1 and T2. Support for invariance over time demonstrates that the construct measured at T2 was similar to the one measured at T1 (MSC was the only multi-item construct measured at T1 and T2). Similarly, there was good support for the cross-level invariance of factor loadings for measures of MSC and the two climate variables (social comparison focus [SCF], and chaos class climate [CCC]). Although not essential, the invariance of factor loadings at the individual student and classroom-level facilitates the interpretation of constructs assessed at both levels by ensuring that they are measured in relation to the same metric. All subsequent results are based upon this invariance model.

Standardized factor loadings are all substantial and statistically significant for multi-item constructs at the individual student level (parameter estimates from the final invariant measurement model are presented in the online supplementary materials, Appendix 3). Thus, for example, self-concept factor loadings vary from .59 to .81 at Time 1 and from .63 to .84 at Time 2. As is typically the case (e.g., Marsh et al., 2009), standardized factor loadings at the L2 classroom level are very high and the residual variance terms are close to zero.

TABLE 1
Correlations Among Constructs

	Correlations Among Individual Student (L1) Constructs															
	SES		*IQ*		*T1MAch*		*T1MSC*		*T2SocCmp*		*T2Chaos*		*T2MACH*		*T2MSC*	
	Est	*SE*	*Est*	*SE*	*Est*	*SE*	*Est*	*SE*	*Est*	*SE*	*Est*	*SE*	*Est*	*SE*	*Est*	*SE*
SES	1.00															
IQ	.18	.03	1.00													
T1MAch_L1	.37	.04	.63	.02	1.00											
T1MSC_L1	.08	.04	.23	.03	.31	.03	1.00									
T2SocCmp_L1	−.07	.03	−.01	.04	−.01	.04	−.03	.03	1.00							
T2Chaos_L1	−.06	.05	.00	.04	.01	.04	−.08	.03	.01	.04	1.00					
T2MAch_L1	.27	.04	.56	.03	.65	.03	.35	.02	−.02	.03	−.04	.04	1.00			
T2MSC_L1	.10	.04	.19	.04	.27	.03	.65	.02	−.05	.03	−.17	.03	.39	.03	1.00	

	Correlations Among Class-Level (L2) Constructs							
	T2Chaos		*T2SocCmp*		*T2MAch*		*T2MSC*	
	Est	*SE*	*Est*	*SE*	*Est*	*SE*	*Est*	*SE*
T2Chaos_L2	1.00							
T2SocCmp_L2	.42	.12	1.00					
T2MAch_L2	−.28	.13	.10	.11	1.00			
T2MSC_L2	−.04	.14	−.54	.13	−.39	.14	1.00	

Note. Parameters more than 1.96 times their standard errors are statistically significant ($p < .05$). (see Figure 1 for a representation of the multilevel structural equation model; and the online supplementary materials, Appendix 3, for the full measurement model, and Appendix 4 for the Mplus syntax). T1 = Time 1; T2 = Time 2; L1 = student level; L2 = class level; MAch = math achievement; MSC = math self-concept; SocCmp = Social-comparison classroom climate; Chaos = chaos classroom climate; est = parameter estimate; SE = standard error.

In CFA/SEM studies, it is often useful to evaluate a CFA measurement model in which all the constructs are merely correlated before testing the final SEM (e.g., Marsh, 2007a; McDonald, 2010). This is also the case for doubly latent ML-SEMs. For example, to better understand the results, it is useful to evaluate the pattern of correlations among the constructs, particularly at the individual student level (Table 1). Thus for example, IQ and T1MAch are substantially correlated ($r = .63$), but T1MSC is more highly correlated with T1MAch ($r = .31$) than IQ ($r = .23$). This supports the domain specificity of the math self-concept and achievement measures. For the two constructs measured at T1 and T2, test-retest correlations are substantial: .65 for MAch and .65 for MSC. Also of note, the correlation between MSC and MAch was higher at T2 ($r = .39$) than T1 ($r = .31$). Consistent with an interpretation of climate constructs as reflecting idiosyncratic noise at the individual student level, all correlations involving these L1 constructs are consistently small (varying from –.07 to +.03).

At the classroom level (Table 1), the two climate variables are positively correlated – classrooms where teachers use more of a focus on social comparison are perceived to have a more chaotic classroom climate ($r = .42$). Chaos is negatively related to MAch ($r = –.28$) but nearly unrelated to MSC. In contrast, a social comparison climate is negatively related to MSC (–.54) but not significantly correlated with MAch. In summary, even though the two classroom climate variables are positively correlated with each other, they have a different pattern of relations with the two outcome variables.

Structural Model: Effects of Climate and Contextual Variables

We now move to the doubly latent ML-SEM that is the focus of this demonstration. Of particular interest are the effects of two classroom climate variables (social comparison focus [T2SCF] and classroom chaos climate [T2CCC]) and the one classroom contextual variable (class-average achievement [T2MAch]). All these climate and contextual variables are based on classroom-level (L2) measures that are aggregates of individual L1 student perceptions within each classroom. The subscript *j* for each of these L2 variables indicates that they take on different values for each classroom but do not vary for students within classrooms. T2 outcome variables are math self-concept (T2MSC) and Math achievement (T2MAch).

For climate effects, the effects are simply the effect of the L2 climate variable on the corresponding L2 outcome variable. Thus, the climate effect is statistically significant if this path coefficient is significantly different from zero. Chaos climate has a statistically significant, negative effect on T2MAch. Students achieve less in classrooms perceived to be chaotic; ES = –.46 (Table 2; also see the online supplementary materials, Appendix 2, for further discussion of the computation of ESs with multilevel data). In contrast to its

TABLE 2
Climate and Contextual Effects: Parameter Estimates (Unstandardized and Standardized) and Effect Sizes

	Unstand		*STDN*		*ES*	
Effect	*Est*	*SE*	*Est*	*SE*	*Est*	*SE*
Climate effects						
L2-Chaos Climate on T2MAch	−.46	.17	−.16	.06	−.46	.17
L2-Chaos Climate on T2MSC	.04	.08	.02	.04	.05	.09
L2-SocCmp Climate on T2MAch	.34	.17	.11	.05	.31	.15
L2-SocCmp Climate on T2MSC	−.23	.11	−.11	.05	−.23	.10
Contextual effect						
L2MAch on T2MSC	−.28	.06	−.16	.04	−.33	.07

Note. Parameters more than 1.96 times their standard errors are statistically significant ($p < .05$). (See Figure 1 for a representation of the multilevel structural equation model; and the online supplementary material, Appendix 4, for the Mplus syntax used to test the model). Unstand = unstandardized parameter estimates; STDN = standardized parameter estimates; ES = effect size; est = parameter estimate; SE = standard error; T2 = Time 2; L2 = class level; MAch = math achievement; MSC = math self-concept; SocCmp = social-comparison classroom climate; Chaos = chaos classroom climate.

effect on achievement, classroom chaos has no significant effect on T2MSC. Students in classes perceived to be chaotic do not have systematically higher or lower MSCs. A classroom climate that emphasizes SCF has a negative predictive effect on T2MSC (ES = –.23). Emphasizing social comparison at the class level is predictive of lower academic self-concepts. In contrast, classrooms characterized by SCF are predictive of higher T2Ach (ES = .31). These results show that the same climate variable can have opposite predictive effects on different desirable outcomes, even though the T2MSC and T2MAch are positively correlated with each other.

For the L2 contextual effects, the effect is defined as the L2 contextual effect on the L2 outcome, minus the effect of the corresponding L1 contextual effect on the L1 outcome.[1] In Figure 1 this difference is represented by the dashed lines leading to the hexagon labeled "contextual effects." Thus, the contextual effect is significantly negative. Consistent with a substantial body of literature, the predictive effect of class-average math ability on math self-concept is negative after controlling for individual achievement (see review by Marsh, Seaton, et al., 2008; also see discussion of substantive basis of the present investigation in the online supplementary materials, Appendix 1). Equally able students tend to have lower academic self-concepts when the class-average achievement level is high.

CRITICAL CONCEPTUAL AND METHODOLOGICAL ISSUES IN CONTEXTUAL AND CLIMATE STUDIES

In this conceptual and methodological demonstration, we found significant effects of classroom context and classroom climate on each of two outcome variables. Consistent with previous research (see review by Marsh, Seaton, et al., 2008; also see the online supplementary materials, Appendix 1), the contextual effect on math self-concept was negative. Classroom chaos climate had a negative predictive effect on MAch but no significant effect on math self-concept. A teacher's focus on social comparison was negatively predictive of math self-concept but positively predictive of MAch. We now consider critical conceptual and methodological issues that have broad generalizability to educational psychology and social science research more generally.

Climate and Contextual Effects Should Be Based on L2 Constructs

It is critical that educational researchers understand that contextual and climate effects should be based on group-level measures—either true L2 constructs or aggregates of L1 measures. In classroom research the group refers to the classroom, but the group could also refer to different levels of analysis (e.g., schools, districts, countries, or even subgroups of students within the same classroom). However, as emphasized earlier, there is an important distinction between contextual and climate variables in the interpretation of the L2 effects in ML-SEMs. For contextual effects, the critical parameter is the L2 effect after controlling for the corresponding L1 effect. For group-mean-centered models like those presented here, this is the difference between the L2 and the L1 effects. Its standard error (to test its statistical significance) and transformations into standardized and ES metrics can then be obtained (see the online supplementary materials, Appendix 4, for the Mplus syntax used in the present investigation).

[1]Contextual effects are the effects of the L2 aggregate after controlling for the effect of the corresponding L1 variable. In multilevel contextual studies, it is typical to use either group-mean or grand-mean centering (see Enders & Tofighi, 2007; Kreft, de Leeuw, & Aiken, 1995; Snijders & Bosker, 1999). For grand-mean centering, the L2 aggregate is confounded with the L1 effects. However, when both the L1 and L2 variables are used to predict the outcome variables, the effect of the L2 variable is controlled for the effect of the corresponding L1 variable. For 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. The appropriate L2 contextual effect can be obtained by subtracting the L1-effect from the L2-effect to disentangle the conflation of L1 effects into L2 estimates due to group-mean centering. When all variables are represented by single indicators (i.e., manifest multilevel models), the choice of group-mean or grand-mean centering is arbitrary, and the results based on one are a simple mathematical transformation of the other. For doubly latent multilevel models (and multilevel CFA/SEM models more generally), the within-group centering facilitates the decomposition of within and between effects and is the basis of all models considered here and is implicitly imposed in all doubly latent ML-SEM models estimated in Mplus. For this reason, to obtain the appropriate L2 contextual effect estimate, it is necessary to subtract L1 effects estimate from L2 effect estimate (for further discussion group and grand mean centering, see the online supplementary materials, Appendix 2).

For climate variables as conceptualized here, the climate effect is the effect of the aggregated L2 construct—not the difference between the L1 and L2 effects. In fact, the main utility of L1 climate ratings is in terms of estimating agreement among students within the class and forming the L2 aggregates; these L1 climate ratings have no substantive meaning in relation to the L2 climate effects. The theoretical and statistical rationale for this claim dates back to the seminal Cronbach (1976) paper on multilevel models but is widely misunderstood in classroom climate research even today. Emphasizing this point, here we found that the residual L1 climate ratings are nearly uncorrelated with other L1 constructs, consistent with an interpretation of residual L1 climate ratings as reflecting idiosyncratic noise at the individual student level.

We leave as an open question whether the residual L1 climate ratings have any substantive role in the interpretation of the results but caution that researchers who make substantive interpretations of residual L1 climate ratings must provide a theoretical and statistical rationale for doing so. This does not mean that the residual L1 climate ratings are meaningless. Indeed, there might well be systematic differences in residual L1 climate ratings by students within classes. For example, there are likely to be systematic method effects (e.g., positive or negative response biases, such that students may give consistently more favorable or less favorable responses in relation to different climate constructs). The position here is not that there are no systematic differences in residual L1 climate responses or that these residual L1 differences necessarily have no meaning, but only that the L2 climate construct must be represented by the L2 aggregation of L1 responses, not by the L1 responses themselves.

To further illustrate this point, consider a study by Papaioannou et al. (2004). They simultaneously evaluated L2 classroom competitiveness in physical education classes (a climate variable) and the L1 competitive orientation of individual students (an individual student difference). For the climate variable, students were asked to rate the climate of the class as a whole (the class was the referent), whereas for the competitive orientation at the L1 individual student level, students were asked to rate their own competitive orientation (the individual student was the referent). In this sense, this study explicitly juxtaposed climate and individual variables, testing a "matching hypothesis" that competitive students are advantaged by being in a class with a more competitive climate (and students low in competitive orientation are advantaged by being in a class with a lower competitive climate). This hypothesis was tested by evaluating the cross-level interaction between individual student (L1) competitive orientation and the class (L2) competitive climate; however, this interaction was not statistically significant. Consistent with that study, if researchers want to study classroom climate, then they should ask students to rate the classroom climate (an L2 class referent) and focus on the L2 aggregations of these L1 perceptions. If researchers want to study individual student differences, then they should ask students to rate themselves (an L1 individual student referent) and focus on these L1 perceptions. If researchers are interested in the juxtaposition of individual student characteristics and corresponding climate variables (as in tests of a matching hypothesis), then researchers should collect ratings of student self-perceptions (with an individual student referent) and ratings of classroom climate (with a classroom referent).

We also note that inappropriate interpretations based on L1 climate ratings in single-level studies might be similar to appropriate interpretations based on L2 climate constructs in multilevel studies. This follows in that the effects of L1 ratings in single-level studies confound the multilevel effects of L1 climate (residuals in the decomposition of L1 ratings in to L1 and L2 constructs) and the critical L2 climate constructs. Thus, for example, in the present investigation there were no significant relations between the L1 residual climate variables and any other constructs. Thus effects of the L1 ratings in a single-level study (that confound the effects of the multilevel L1 and L2 effects and are inappropriate as a climate construct) would coincidentally result in a similar interpretation as the effects of the L2 climate measure in the more appropriate multilevel model. However, the single-level model would still be inappropriate. From a statistical perspective, the nonindependence of responses by students within the same classroom would typically inflate Type 1 error rates substantially in single-level studies. More important, this fortuitous coincidence of interpretations based on appropriate and inappropriate models can only be evaluated if the appropriate model is evaluated—in which case the inappropriate single-level model would unlikely to be pursued. Nevertheless, we suspect that this phenomenon might explain why so many applied researchers have continued to use single-level models even though their inappropriateness has been well known for more than a quarter of a century.

Multilevel models considered here allow intercepts to vary (random intercept models) but do not allow slopes to vary across different L2 classrooms. An obvious extension of this model is to allow slopes to vary or to include cross-level interactions between L1 and L2 constructs. Using the BFLPE as a substantive example, Marsh et al. (2009) demonstrated an extension to the doubly latent model to include latent interactions and nonlinear effects. Furthermore, Preacher and colleagues (Preacher, Zhang, & Zyphur, 2011; Preacher, Zyphur, & Zhang, 2010) have built on our earlier work to develop a general multilevel SEM framework for assessing multilevel mediation when estimating the effect of group-level variables.

Sampling Error in Contextual and Climate Variables

In the doubly latent ML-SEM, (Lüdtke et al., 2008; Lüdtke et al., 2011; Marsh et al., 2009), the observed class-average is assumed to be a fallible measure of the true class mean.

Hence, the observed class mean is corrected for sampling error so that analyses are based on a latent class mean (i.e., latent in relation to sampling error). For classroom climate measures we argue that it is always appropriate to test for sampling error and that the doubly latent model is appropriate so long as there are multiple indicators of each construct. In particular, there is potential measurement error at L1 and L2 due to sampling of items, and there is sampling error in going from L1 to L2 based on sampling of persons.

For classroom contextual measures like class-average achievement context, there is some ambiguity in the appropriate operationalization in the doubly latent model in relation to sampling error in the model presented here (e.g., Figure 1). If each class is considered as a sample of students from a larger population of students, then it is appropriate to control for sampling error and this is done automatically in our doubly latent ML-SEM. As noted earlier, estimates of sampling error are smaller when agreement among students within the same class is higher and the number of students with the class is larger. We argue that this is always appropriate when achievement is a dependent variable (e.g., used as an outcome to evaluate the effects of classroom climate) and is the standard approach in multilevel modeling. However, for contextual variables such as class-average achievement, it can be argued that there is no sampling error—at least not when all students in each class are sampled (i.e., a sampling ratio approaches 1.0). In this case it might be more appropriate for L2 constructs to be represented by manifest variables or indicators such that sampling error is assumed to be zero. Hence, by controlling for sampling error in the doubly latent ML-SEM in contextual models, the applied researcher might be overcorrecting for unreliability due to sampling error. Marsh et al. (2009; Lüdtke et al., 2011; Lüdtke et al., 2008; also see Skrondal & Laake, 2001)[2] argued that the "best" estimate of the contextual effect typically lies somewhere between the doubly latent model (that might overcorrect for sampling error, leading to inflated estimates of the contextual effect) and an alternative model that fixed sampling error to zero (that might undercorrect for sampling error, leading to underestimates of the size of the contextual effect). In supplemental analyses based on this alternative model for data considered here, we evaluated the contextual effects when sampling error was constrained to be zero (also see Marsh et al., 2009). Consistent with expectations, the size of the negative effect of class-average achievement on math self-concept was marginally smaller. However, because the ICC2 for achievement was extremely high (.957, indicating that there was almost no sampling error even when it was included in the doubly latent ML-SEM), the implications for controlling or not controlling for sampling error were trivial in our empirical example.

Marsh et al. (2009) argued that when the sampling ratio is small (e.g., only a small proportion of the students in each group are sampled), it is appropriate to control for sampling error. Others (e.g., Hutchison, 2007; also see Shin & Raudenbush, 2010) argue that a class of students should always be considered a sample of a larger superpopulation of potential students so that it is typically appropriate to control for sampling error. We clearly agree with this contention in relation to climate variables where the traditional MLM approach in the doubly latent model is appropriate. However, for contextual variables the decision of whether to control for sampling error depends in part on the nature of the research question and the sampling ratio. In particular, if the focus of the study is on the context of each particular classroom and the class-average contextual variable is based on all students within the class, then it is reasonable to argue that the class-average measure is a population value that contains no sampling error. Although the students in a class represent a sample from a larger population of students, the contextual effect is based on the average value of students actually in the class, not some hypothetical group of students who might have been in the class. However, when class-average ability is an L2 outcome variable rather than an L2 contextual variable, it is appropriate to treat it as a sample measure with sampling error. That is, in this case students within the class only form a subsample from the potentially larger sample of students that could have formed the class and to which we want the results to generalize. In either case it is appropriate to control the L2 aggregated variable for measurement error due to sampling of items based on multiple indicators of the L2 construct.

In summary, the appropriate control for sampling error in contextual models has not been fully resolved so that it is incumbent upon applied researchers to defend their operationalization, particularly when estimates of sampling error are large (i.e., when ICC1 and ICC2 are small) or the sampling ratio is small (i.e., aggregated contextual measures are based on a small subsample of students from each class).

Assumptions of Causality and Underlying Processes

Classroom climate and classroom contextual studies are typically based on correlational analyses so that causal interpretations should be offered tentatively and interpreted cautiously. Here, as with all social science research, it is appropriate to hypothesize causal relations, but researchers should fully interrogate support for causal hypotheses in relation to a construct validity approach (see Marsh, 2007b) based on multiple indicators, multiple (mixed) methods, multiple experimental

[2]Different research disciplines have used different terminology relevant to contextual and climate variables. Lüdtke and colleagues (Lüdtke et al., 2011; Lüdtke et al., 2008; Marsh et al., 2009; also see Skrondal & Laake, 2001) described the distinction between *formative* and *reflective* aggregations of L1 variables that is closely related to our distinction between contextual and climate variables. These terms come from a factor analysis perspective. However, a similar distinction is made in organisational psychology (e.g., Bliese, 2000; Bliese et al., 2007; also see Kozlowski & Klein, 2000) between compilation (or configural) models and composition models.

designs, and multiple time points, as well as testing the generalizability of the results across diverse settings and measures. The evidence for construct validity includes the content, response processes by participants, internal structure in terms of consistency and factor structure, and convergent and discriminant validity in relation to other constructs—including, for example, experimental and quasi-experimental manipulations, criterion-related validity, and validity generalisation to relevant and similar situations or populations. Particularly for contextual and climate measures based on aggregates of student-level responses, it is sometimes possible to have direct measures of the group-level construct that do not depend on the aggregation responses by individuals (classroom observations based on responses by external observers; see Lüdtke et al., 2005). Although stronger inferences about causality are possible in longitudinal, quasi-experimental, and true experimental (with random assignment) studies, trying to "prove" causality is usually a precarious undertaking, based on typically implicit assumptions that are untested or untestable. Even in true experimental studies in applied social science disciplines, there is typically some ambiguity regarding the interpretation of what was actually manipulated, how it varies with different subgroups within the population, and its relevance to theory and typical practice. The problems of causal interpretations with contextual studies have been discussed extensively in the organizational psychology (Bliese et al., 2007) and in the social sciences more generally (e.g., Foster, 2010; Morgan & Winship, 2007).

Support for the validity interpretations of potential causal relations of climate and contextual effects are unlikely to be sufficient based on a single study but might be stronger when based on cumulative results from an ongoing research program. To illustrate the type of support that might be developed in support of construct validity interpretations, we briefly summarize some of the main findings in support for the interpretation of the negative effect of class- or school-average ability on academic self-concept (results summarized here are presented in more detail in review by Marsh, 2007b; Marsh et al., 2008). The pattern of results is similar for studies based on measures of class-average and school-average achievement. Quasi-experimental, longitudinal studies based on matching designs as well as statistical controls, show that academic self-concept declines when students shift from mixed-ability schools to academically selective schools over time (based on pre–post comparisons) and in relation to students matched on academic ability who continue to attend mixed-ability schools. Extended longitudinal studies show that the size of this negative effect of school-average ability grows more negative the longer students attend a selective school and is maintained even 2 and 4 years after graduation from high school. Also, there is good support for the convergent and discriminant validity of the negative effect of school-average ability as it is largely limited to academic components of self-concept and nearly unrelated to nonacademic components of self-concept and to self-esteem. Studies based on Organisation for Economic Co-operation and Development–Programme for International Student Assessment data from many different countries shows that the negative effect of school-average ability has good cross-national generalizability (see review by Marsh, Seaton, et al., 2008). Although the "third variable" problem is always a threat to contextual studies that do not involve random assignment, Marsh et al. (2008) argued that this is an unlikely counter-explanation of the results. In particular, most potential "third variables" (resources, per student expenditures, SES, teacher qualifications, etc.) tend to be positively related to school-average achievement, so that controlling for them would increase the size of the negative effect of school-average achievement on academic self-concept. As illustrated with the interpretation of this effect, the construct validity of interpretations of contextual and climate effects should be based on accumulated evidence over a variety of different studies rather than the results of a single study.

SUMMARY AND CONCLUSIONS

Educational research is inherently multilevel, but there is a substantial research literature into the effects of classroom contexts and climates that has either ignored this multilevel perspective or apparently misunderstood it. Even when assessment of classroom climates and contexts are based upon responses by individual students, it is crucial that the contextual and climate effects are based on class-average aggregates of the individual student responses. Many studies of classroom climate and context have inappropriately based interpretations on the L1 constructs. Particularly the results of the school-average ability on academic self-concept provide a classic example of the inappropriateness of not clearly distinguishing L1 and L2 effects. At the individual student level (L1), achievement is positively related to academic self-concept (the brighter I am, the better my academic self-concept). However, at the class or school (L2) level, school- or class-average achievement is negatively related to academic self-concept (the brighter everyone else in my class or school, the lower my academic self-concept). Inferences about the effects of classroom climate or context must be based on classroom-level measures, either true classroom-level variables measured at the level of the classroom or aggregations of scores by individual students to form classroom-level aggregates. Ratings by students within each class should not be used to infer classroom climates.

Historically, the dominant approaches in the evaluation of educational data have been CFA and SEM based on multiple indicators and MLM based on multiple levels. CFA and SEM approaches have been used traditionally in single-level analyses to assess support for a priori factor structures, and to assess and control for measurement error. MLMs have traditionally been used with hierarchical data to correct tests of statistical significance due to the inherent violations of

assumptions of independence in multilevel data, and to unconfound the effects of variables at different levels of analysis. Both these approaches are critical for the appropriate analysis of climate and contextual effects. However, it is only recently that there has been an integration of these approaches in a form that is practically useful for the applied researcher. Here we demonstrated the application of the doubly latent ML-SEM—an integration of CFA, SEM, and MLMs—to evaluate classroom contextual effects and its extension to the analysis of classroom climate effects. The model is doubly latent—latent in the sense of controlling for measurement error at L1 and L2 due to sampling of items, and latent in the sense of controlling for sampling error in the aggregation from L1 to L2 due to sampling of students. Like traditional applications of CFA in single-level studies, the doubly latent approach allows researchers to test support for the a priori factor structure at L1. However, it also allows researchers to test a priori factor structures at L2 and the invariance of the factor structure across the L1 and L2 levels. Because this approach has not been widely used in applied research, much work is still needed to establish best practice and appropriate limitations—particularly when sample sizes at L1 or L2 are modest (see Lüdtke et al., 2011; Lüdtke et al., 2008). Nevertheless, the doubly latent ML-SEM offers exciting new potential for the evaluation of classroom climates and contexts that has broad applicability to organizational research seeking to juxtapose the effects of individuals and the groups to which they belong.

GUIDELINES FOR ANALYSIS OF CLASSROOM CLIMATE (AND CONTEXT) BASED ON STUDENT REPONSES

As noted earlier, one purpose of this article is to provide guidelines for appropriate analysis of particularly classroom climate studies that use aggregates of student-level responses to infer classroom constructs. For this reason we conclude with a summary of the present investigation in the form of a set of guidelines for classroom climate and contextual studies. Although many of these guidelines are relevant to both classroom climate and contextual studies, the methodological problems are more widespread in classroom climate research that is the main focus of this article and extension of our doubly latent ML-SEM. Although we focus on classroom research in educational settings, the issues are relevant to all disciplines that use aggregates of L1 individual responses to represent an L2 group-level construct.

1. Interpret classroom climate and contextual constructs at the classroom level. The interpretation of L1 student-level responses as if it were an L2 classroom-level construct is not appropriate. This is clearly the most serious concern and a major focus of our article.
2. Design classroom climate measures such that the referent is the L2 classroom, not the individual student or a mixture of individual student and classroom referents. Thus, for example, it is quite reasonable to assess the competitive and cooperative orientations of individual students and relate these to other variables. However, it is wrong to assume that these individual difference variables reflect classroom climate. Although classroom climate might be related to these variables, it is important to keep them separate.
3. In multilevel studies, control for measurement error at both the L1 and L2 levels through the use of multiple indicators. It is well known that the failure to control for L1 measurement error can lead to systematic biases in the interpretation of L2 constructs. For example, Harker and Tymms (2004) noted that based on highly reliable pretests of individual student achievement, there was no contextual effect of school-average achievement on subsequent achievement. However, as they added more and more measurement error to their pretest achievement measure, the contextual effect of achievement became increasingly positive. In that study the apparent contextual effect was completely an artifact of the failure to control for measurement error—an effect labeled as the phantom effect. However, there is measurement error at both the L1 individual student and the L2 classroom levels that should be controlled by multiple indicators of each L1 and L2 construct. Measurement error at L2 attenuates the effects of L2 classroom constructs in the same way the measurement error at L1 attenuates the effects of L1 variables on L1 outcomes. In both classroom climate and contextual studies considered here, the multiple indicators of L2 constructs are aggregates of the multiple indicators of the L1 constructs.
4. Evaluate the ICC1 and ICC2 of classroom climate measures. If the ICC1 for a climate variable is statistically nonsignificant or close to zero, the appropriate interpretation is that there is little or no class-to-class variation in the scores used to represent the climate. Although this could mean that either there really are no differences or the measures being used are not appropriate, it makes little sense to pursue further analyses based on these measures. If the number of students sampled from each class is small so that ICC2 is small, it is particularly important to control for sampling error using appropriate multilevel models like those demonstrated here.
5. Base classroom climate and contextual studies on a large number of classes. From a multilevel perspective, the effective sample size is the number of classes, not just the number of students. Particularly the doubly latent ML-SEM model used here requires a large number of classes. Although there are no golden rules, Lüdtke and colleagues (Lüdtke et al., 2011; Lüdtke

et al., 2008) suggested that at least 50 classes are needed; even more classes are preferable, particularly when ICC1 and ICC2 are modest. However, they also described submodels that might be appropriate when the number of classes is small. Nevertheless, even here there is a complicated trade-off between systematic bias associated with submodels and sampling error associated with the doubly latent model when the number of classes is modest. Based on these recommendations, it appears that the majority of classroom climate and contextual studies do not consider a sufficient number of classes.

6. Test the latent factor structure at both the student and classroom level. Do not assume that the factor structure relating multiple indicators to latent constructs is necessarily the same at the individual student and classroom level. Although classroom researchers sometimes evaluate the structure at L1, it is often the structure at L2 that is more relevant. Even if responses have a well-defined structure at L1, there is no guarantee that the structure is well defined at L2. Thus, for example, math and verbal achievement are moderately correlated at the individual student level (e.g., .5 to .7) but tend to be so highly correlated at the class level as to be almost indistinguishable (or appropriately represented as a single global achievement factor). For classroom climate studies it might only be the L2 factor structure that is relevant. For classroom contextual studies, both the L1 and L2 structures are critical, and interpretations are facilitated if the factor structures are invariant over L1 and L2. However, guidelines for the evaluation of the fit of multilevel models and tests of cross-level invariance have not been resolved.
7. Prior to evaluating doubly latent ML-SEMs like those considered here, evaluate doubly latent ML-CFA measurement models. These more basic models should be evaluated in relation to goodness of fit and to construct validity based on an evaluation of parameter estimates and relations among constructs. Models designed to test more complicated hypotheses of multilevel mediation and moderation should begin with more basic ML-SEMs like those demonstrated here.
8. Be appropriately cautious in the interpretation of correlational effects as causal effects, particularly in cross-sectional studies based on single wave of data.

ACKNOWLEDGMENTS

This research was supported in part by grants to the first author from the UK Economic and Social Research Council and from the King Saud University in Saudi Arabia, and a grant of the state of Baden-Wuerttemberg (Germany) to the fourth author.

REFERENCES

Ames, C. (1992). Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology*, *84*, 261–271.

Baumert, J., Lehmann, R. H., Lehrke, M., Schmitz, B., Clausen, M., Hosenfeld, I., … Neubrand, J. (1997). *TIMSS: Mathematisch-Naturwissenschaftlicher Unterricht im internationalen Vergleich.* [TIMSS: Mathematics and science instruction in an international comparison]. Opladen, Germany: Leske and Budrich.

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.

Bliese, P. D., Chan, D., & Ployhart, R. E. (2007). Multilevel methods: Future directions in measurement, longitudinal analyses, and nonnormal outcomes. *Organizational Research Methods*, *10*, 551–563.

Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer-Verlag.

Brophy, J. (1988). Educating teachers about managing classrooms and students. *Teaching and Teacher Education*, *4*, 1–18.

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.

Covington, M. V. (2001). Goal theory, motivation, and school achievement: An integrative review. *Annual Review of Psychology*, *51*, 171–200.

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

Croon, M. A., & van Veldhoven, M. J. P. M. (2007). Predicting group level variables from variables measured at the individual level: A latent variable multilevel model. *Psychological Methods*, *12*, 45–57.

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.

Evertson, C., Emmer, E., Sanford, J., & Clements, B. (1983). Improving classroom management: An experiment in elementary school classrooms. *The Elementary School Journal*, *84*, 173–188.

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.

Foster, E. M. (2010). Causal inference and developmental psychology. *Developmental Psychology*, *46*, 1454–1480.

Goldstein, H. (2003). *Multilevel statistical models*. London, UK: Edward Arnold.

Goldstein, H., & McDonald, R. P. (1988). A general-model for the analysis of multilevel data. *Psychometrika*, *53*, 455–467.

Harker, R., & Tymms, P. (2004). The effects of student composition on school outcomes. *School Effectiveness and School Improvement*, *15*, 177–199.

Heckhausen, J., & Heckhausen, H. (2008). *Motivation and action* (2nd ed.). New York, NY: Cambridge University Press.

Helmke, A., Schneider, W., & Weinert, F. E. (1986). Quality of instruction and classroom learning outcomes: The German contribution to the IEA Classroom Environment Study. *Teaching & Teacher Education*, *2*, 1–18.

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods*, *5*, 64–85.

Hutchison, D. (2007). When is a compositional effect not a compositional effect? *Quality & Quantity*, *41*, 219–232.

Iverson, G. R. (1991). Contextual analysis. *Sage University Paper Series on Quantitative Approaches in the Social Sciences*, 07–081. Newbury Park, CA: Sage.

Jöreskog, K. G., & Sörbom, D. (1988). *LISREL 7 - A guide to the program and applications* (2nd ed.). Chicago, IL: SPSS.

Kane, M. T., Gillmore, G. M., & Crooks, T. J. (1976). Student evaluations of teaching: The generalizability of class means. *Journal of Educational Measurement*, *13*, 171–183.

Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions*. Newbury Park, CA: Sage.

Kozlowski, S. W. J., & Klein, K. J. (2000). A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes. In K. J. Klein & S.W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 3–90). San Francisco, CA: Jossey-Bass.

Kreft, I. G. G., de Leeuw, J., & Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research*, *30*, 1–21.

LaHuis, D. M., & Ferguson, M. W. (2009). The accuracy of significance tests for slope variance components in multilevel random coefficient models. *Organizational Research Methods*, *12*, 418–443.

Lewis, R. (2001). Classroom discipline and student responsibility: The students' view. *Teaching & Teacher Education*, *17*, 307–319.

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

Lüdtke, O., Köller, O., Marsh, H. W., & Trautwein, U. (2005). Teacher frame of reference and the big-fish-little-pond effect. *Contemporary Educational Psychology*, *30*, 263–285.

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

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

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

Mainhard, M. T., Brekelmans, M., & Wubbels, T. (2011). Coercive and supportive teacher behaviour: Within- and across-lesson associations with the classroom social climate. *Learning and Instruction*, *21*, 345–354.

Marsh, H. W. (1990). Causal ordering of academic self-concept and academic achievement: A multiwave, longitudinal panel analysis. *Journal of Educational Psychology*, *82*, 646–656.

Marsh, H. W. (2007a). 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 ed., pp. 774–798). Hoboken, NJ: Wiley.

Marsh, H. W. (2007b). *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., Byrne, B. M., & Yeung, A. S. (1999). Causal ordering of academic self-concept and achievement: Reanalysis of a pioneering study and revised recommendations. *Educational Psychologist*, *34*, 155–167.

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., 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., Martin, A. J., & Cheng, J. H. S. (2008). A multilevel perspective on gender in classroom motivation and climate: Potential benefits of male teachers for boys? *Journal of Educational Psychology*, *100*, 78–95.

Marsh, H. W., Seaton, M., Trautwein, U., Lüdtke, O., Hau, K. T., O'Mara, A. J., & Craven, R. G. (2008). The big-fish-little-pond-effect stands up to critical scrutiny: Implications for theory, methodology, and future research. *Educational Psychology Review*, *20*, 319–350.

Marshall, H. H., & Weinstein, R. S. (1984). Classroom factors affecting students' self-evaluations. *Review of Educational Research*, *54*, 301–326.

Matheny, K. B., & Edwards, C. R. (1974). Academic improvement through an experimental classroom management system. *Journal of School Psychology*, *12*, 222–232.

McDonald, R. P. (1993). A general-model for 2-level data with responses missing at random. *Psychometrika 58*, 575–585.

McDonald, R. P. (1994). The bilevel reticular action model for path-analysis with latent-variables. *Sociological Methods & Research*, *22*, 399–413.

McDonald, R. P. (2010). Structural models and the art of approximation. *Perspectives on Psychological Science*, *5*, 675–686.

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

Midgley C. 2002. *Goals, goal structures, and patterns of adaptive learning*. Hillsdale, NJ: Erlbaum.

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

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.

Mischo, C., & Rheinberg, F. (1995). Erziehungsziele von Lehrern und individuelle Bezugsnormen der Leistungsbewertung [Educational goals and teachers' preference of individual reference-norms in evaluating academic achievement]. *Zeitschrift für Pädagogische Psychologie*, *9*, 139–151.

Möller, J., & Köller, O. (2001). Frame of reference effects following the announcement of exam results. *Contemporary Educational Psychology*, *26*, 277–287.

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference: Methods and principles for social research.* Cambridge, UK: Cambridge University Press.

Muthén, B. (1989). Latent variable modelling in heterogeneous populations. *Psychometrika*, *54*, 557–585.

Muthén, B. O. (1994). Multilevel covariance structure-analysis. *Sociological Methods & Research*, *22*, 376–398.

Muthén, L. K., & Muthén, B. O. (2006–2010). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.

Nagengast, B., & Marsh, H. W. (2011). The negative effect of school average ability on science self-concept in the UK, the UK countries and the world: The Big-Fish-Little-Pond-Effect for PISA 2006, *Educational Psychology*, *31*, 629–656.

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 advantage 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). Generalized multilevel structural equation modeling. *Psychometrika*, *69*, 167–190.

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

Rheinberg, F. (1980). *Leistungsbewertung und Lernmotivation* [Performance evaluation and motivation for learning]. Göttingen, Germany: Hogrefe.

Rheinberg, F., & Krug, S. (1999). *Motivationsförderung im Schulalltag* [Enhancing motivation in school]. Göttingen, Germany: Hogrefe.

Rutter, M., Maughan, B., Mortimore, P., Ouston, J., & Smith, A. (1979). *Fifteen thousand hours: Secondary schools and their effects on children*. Cambridge, MA: Harvard University Press.

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

Shin, Y., & Raudenbush, S. W. (2010). A latent cluster-mean approach to the contextual effects model with missing data. *Journal of Educational and Behavioral Statistics*, *35*, 26–53.

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

Snijders, T. A. B., Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.

Trautwein, U., Lüdtke, O., Köller, O. & Baumert, J. (2006). Self-esteem, academic self-concept, and achievement: How the learning environment moderates the dynamics of self-concept. *Journal of Personality and Social Psychology*, *90*, 334–349.

Trautwein, U., Lüdtke, O., Marsh, H. W., Köller, O., & Baumert, J. (2006). Tracking, grading, and student motivation: Using group composition and status to predict self-concept and interest in ninth grade mathematics. *Journal of Educational Psychology*, *98*, 788–806.

Wang, M.-T., & Holcombe, R. (2010). Adolescents' perceptions of school environment, engagement, and academic achievement in middle school. *American Educational Research Journal*, *47*, 633–662.

Online Supplemental Materials for
Classroom Climate and Contextual Effects:
Conceptual and Methodological Issues in the Evaluation of Group-level Effects

Appendix 1: Substantive Basis of the Present Investigation

Reciprocal Effects Between Academic Self-concept and Achievement

The causal ordering of academic achievement (ACH) and academic self-concept (ASC) has inspired a long debate, competing theoretical models, and many empirical studies (Marsh, 1990; Marsh & Craven, 2006). Calsyn and Kenny (1977) contrasted the skill-development model (ACH ➔ ASC) with the self-enhancement model (ASC➔ACH). Integrating these two models, the reciprocal effects model (Marsh & Craven, 2006; Marsh & O'Mara, 2008; Marsh, Trautwein, Lüdtke, Köller & Baumert, 2005) predicts that ASC and achievement are reciprocally related—each is a cause and an effect of the other. Empirical tests of the reciprocal effects model are based on multivariable panel studies in which ASC and achievement are measured on two or more occasions. In their meta-analysis, Valentine, DuBois, and Cooper (2004) concluded that there was clear support for the predictions of the reciprocal effects model. They found positive self-beliefs to predict later academic achievement, even when initial levels of achievement were controlled. Longitudinal research demonstrates that these effects are long-lasting (e.g., Guay, Marsh, & Boivin, 2003; Marsh & O'Mara, 2008). A critical feature of this research is on the need to evaluate ACH and ASC on at least two occasions, emphasizing dual roles of ASC as both a valued outcome and a facilitator of desirable outcomes.

Big-Fish-Little-Pond-Effect (BFLPE): A Contextual Effect

The big-fish-little-pond effect (BFLPE) is a widely studied contextual effect in educational research (Marsh, 2007b; Marsh & Craven, 2006; Marsh, Seaton, et al., 2008). The BFLPE is a classic contextual effect in which the effect of individual student achievement (L1-ACH) on academic self-concept (L1-ASC) is positive, but the corresponding effect of group-average (school or classroom) achievement (L2-ACH) is negative. Students evaluate their accomplishments in relation to those of their classmates and use this comparative self-evaluation to determine their ASCs. The BFLPE is a robust, long-lasting contextual effect that generalises across diverse research settings, levels of education, and cultures from all over the world (Marsh et al., 2008). From a policy perspective, the BFLPE provides an alternative, contradictory perspective to educational policy on the placement of students in special education settings, one that is being enacted in many countries throughout the world. The juxtaposition of the positive effects of individual achievement and the negative effects of class-average achievement is inherently a multilevel issue that cannot be represented adequately at either the individual or the classroom level. Hence, it is important to analyze data with appropriate multi-level statistical procedures. Demonstrating the synergy between applied research and methodology, methodological limitations in existing BFLPE research led to the application and extension of the doubly-latent multilevel model used here (Marsh, Lüdtke, et al., 2009; Marsh et al., 2008).

Social Comparison Feedback (SCF) and Classroom Chaos: Climate effects

Particularly in German research there is a substantial literature on classroom climate that focuses on the nature of student feedback given by teachers to individual students (e.g., Heckhausen & Heckhausen, 2008; Lüdtke, Köller, Marsh & Trautwein, 2005; Rheinberg & Krug, 1999; also see related work by Ames, 1992; Covington, 2001; Marshall & Weinstein, 1984). Rheinberg (1980) distinguished teachers who prefer a social-comparison standard from teachers who prefer an individual standard. Teachers using an individualized standard provided temporal feedback to students and emphasized improvement, whereas teachers with a social-comparison frame of reference assessed their students' accomplishments on the basis of comparisons with others. The central characteristic of an individual frame of reference when assessing students' accomplishments is the emphasis on the intra-individual improvement of individual students, thereby enhancing students' self-concept, motivation, and achievement. Subsequently, several studies—including correlational, experimental, and longitudinal designs—have indeed found positive effects of an individualized frame of reference on students' attitudes, attributions, and motivation (e.g., Heckhausen & Heckhausen, 2008; Mischo & Rheinberg, 1995; Rheinberg, 1980; Rheinberg & Krug, 1999). More specifically, Lüdtke et al. (2005) demonstrated that an individualized teacher frame of reference enhanced ASC.

Classroom management is a basic function commonly performed by teachers that influences instructional activities and student outcomes (Brophy, 1988; Evertson, Emmer, Sanford & Clements, 1983; Lewis, 2001; Matheny & Edwards, 1974). Thus Brophy emphasized the importance of the "teacher's ability to maximize the time that students spend actively engaged in worthwhile academic activities, to minimize the time that they spend waiting for activities to get started, making transitions between activities, sitting

with nothing to do, or engaging in misconduct" (Brophy, 1988, p. 3). Rutter, Maughan, Mortimore, Ouston, and Smith (1979) emphasized that more organized teachers, who did not waste time in transitory periods (e.g., handing out papers) had better behaved students. Helmke, Schneider, and Weinert (1986) reported that efficient use of time was positively correlated with student achievement. In summary, chaotic classroom climates are likely to be antithetical to student behavior and outcomes.

References (also see References in the Published Article)

Ames, C. (1992). Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology, 84,* 261-271

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.

Brophy, J. (1988). Educating teachers about managing classrooms and students. *Teaching and Teacher Education, 4,* 1-18.

Calsyn, R. J., & Kenny, D. A. (1977). Self-concept of ability and perceived evaluation of others: Cause or effect of academic achievement? Journal of Educational Psychology, 69, 136-145.

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

Covington, M. V. (2001). Goal theory, motivation, and school achievement: An integrative review. *Annual Review of Psychology, 51,* 171-200.

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.

Evertson, C., Emmer, E., Sanford, J., & Clements, B. (1983). Improving classroom management: An experiment in elementary school classrooms. *The Elementary School Journal, 84*, 173-188.

Guay, F., Marsh, H. W., & Boivin, M. (2003). Academic self-concept and academic achievement: Developmental perspectives on their causal ordering. *Journal of Educational Psychology, 95*, 124-136.

Heller, K. A., Gaedike, A.-K.., & Weinlader, H. (1985). *Kognitiver Fahigkeits-Test* (KFT). Weinheim: Beltz.

Helmke, A., Schneider, W., & Weinert, F. E. (1986). Quality of instruction and classroom learning outcomes: The German contribution to the IEA Classroom Environment Study. *Teaching & Teacher Education, 2*, 1– 18.

Jerusalem, M. (1984). Reference group, learning environment and self-evaluations: A dynamic multi-level analysis with latent variables. In R. Schwarzer (Ed.), The self in anxiety, stress and depression (pp. 61-73). North-Holland: Elsevier Science Publishers.

Marsh, H. W., Hau, K.-T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research, 33*, 181–220.

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., 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.

Martin, M. O., & Kelly, D. L. (Eds.). (1997). Third international mathematics and science study. Technical report: Vo.. II. *Implementation and analysis. Primary and middle school years.* Chesnut Hill, MA: Boston College.

O'Brien, R. M. (1990). Estimating the reliability of aggregate-level: Variables based on individual-level characteristics. *Sociological Methods and Research, 18,* 473-504.

Raykov, T., & Marcoulides, G. A. (2004). Using the Delta Method for Approximate Interval Estimation of Parametric Functions in Covariance Structure Models. *Structural Equation Modeling, 11,* 659-675.

Searle, S. R., Casella, G.,&McCulloch, C. E. (1992). Variance components. New York, NY: Wiley & Sons.

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.

Online Supplemental Materials:
Appendix 2: More detailed description of the Methodology

The present investigation is based on the extended German sample from the Third International Mathematics and Science Study (TIMSS) and is a reanalysis and extension the Lüdtke, Marsh, et al. study (2005) where the sample and procedures are described in detail. The sample is nationally representative with respect to region, school type, and gender, based on achievement and psychological variables for 2,261 students in 128 classes collected in Grades 7 and 8. Nearly all (95%) participants were German. There was complete data for all measures of IQ and achievement, and nearly complete data for all Time 1 psychological constructs. However, because there was considerable missing data for the T2 psychological constructs (% non-missing varied 77% to 79%), we implemented the full-information-maximum-likelihood approach to missing data as implemented in Mplus (Muthén & Muthén, 2006-2010; also see Little & Rubin, 1987; Schafer, 1997).

Math achievement in Grade 8 (T2) was part of the official TIMSS (see Martin & Kelly, 1997) whilst the 36 math items in Grade 7 (T1) were taken from previous *International Association for the Evaluation of Educational Achievement* studies (see Baumert, Roeder, Sang, & Schmitz, 1986). Coefficient alpha estimates of reliability were .81 (T1) and .88 (T2). Intelligence (IQ, T1 only) was measured by the Kognitiver Fahigkeitstest (KFT; Heller, Gaedike & Weinlader, 1985). The KFT (Cognitive Abilities Test) is a composite of verbal, numerical and nonverbal/figural abilities. There were two indicators of socioeconomic status based on mother's and father's education. Math self-concept was assessed with a short 4-item scale (e.g., "Some topics in math are just so hard that I know from the start I'll never understand them") that has been used widely in many large-scale German studies with demonstrated reliability and validity (e.g., Möller & Köller, 2001). Coefficient alpha estimates of reliability were .84 (T1) and .85 (T2). Students assessed their math teacher's social comparison feedback at the end of Grade 8 using four items developed by Jerusalem (1984): e.g., "If a student improves his/her achievement, the teacher praises him/her, even if he/she is below class average." High scores on this scale indicate that the teacher was perceived to have a pronounced social comparison frame of reference that is typical in Germany (as opposed to an individualized frame of reference). The coefficient alpha estimate of reliability was .85 (T2 only). The Classroom Chaos Climate (Chaos) measure was based on responses to four items (e.g., "A lot of time is wasted during the lesson"; "In the beginning of each lesson it takes a long time before students are quiet"; see Fend & Specht, 1986) and had a coefficient alpha estimated reliability of .85 (T2 only).

Statistical analysis

All statistical analyses were done with Mplus (version 6; Muthén & Muthén, 2006-2010) using the doubly-latent MLSEM summarized earlier and described in greater detail by Marsh, Lüdtke, et al. 2009; Lüdtke et al., 2008; 2011; also see Appendix 4 for Mplus syntax). In order to facilitate interpretation of the coefficients in multilevel models and to reduce multicollinearity all variables were standardized (M = 0, SD = 1). 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, 2006-2010) as well as the robust $\chi 2$ test statistic and inspection of parameter estimates. However, we also compared the relative fit of different models in a nested or partially nested taxonomy of models designed a priori to evaluate particular aspects of interest than single models (Marsh, 2007a; Marsh, et al., 2009). Nevertheless, we emphasize that traditional cut-off values for single models and the comparison of different models only constitute rough guidelines (Marsh, 2007a; Marsh et al., 2005; also see Marsh, Hau, Balla & Grayson, 1998).

Interpretation of Classroom contextual and climate effects: Standardization and Effect Sizes

Contextual effects. For aggregated contextual variables (like class-average achievement in the present investigation), the critical parameter is the L2 effect after controlling for the effects of interindividual (L1) differences. Importantly, the L1 variable (e.g., individual student achievement) that was aggregated to form the L2 contextual variable is meaningful in its own right as a potentially important individual difference variable, and is typically included in the contextual model. The L1 measure (individual student achievement) and the corresponding L2 contextual variable (class-average achievement) are different variables. In MLMs it is typical to distinguish between group-mean centering and grand mean centering (Enders & Tofighi, 2007; Kreft, de Leeuw & Aiken, 1995; Lüdtke, Robitzsch, Trautwein, & Kunter (2009). In all models considered here, we used grand-mean centering for all L1 constructs measured only at the student level (e.g., SES and IQ). For L1 variables aggregated to form L2 variables, there is an implicit group-mean centering in the decomposition of variables within (L1) and between (L2) effects (i.e. the latent mean of class *j* is subtracted from the score of student *i* in class *j)*. Hence, all variables that appear at both L1 and L2 in the doubly-latent models are implicitly group-mean centered. In this case, a contextual effect is present if the L2 (between-

class) regression coefficient is significantly different from the L1 (within-class regression) coefficient. A test of this hypothesis was accomplished by calculating an additional parameter that provides a direct estimate of the contextual effect (Enders & Tofighi, 2007; Kreft, et al, 1995; see Supplemental Materials, Appendix 4 for Mplus syntax): the difference between the regression weights assigned to L1 and L2-ACH. Although this subtraction could easily be done by simple inspection, the important advantage of the approach used here is that it is easily transformed into an effect size (ES) metric and provides standard errors obtained with the multivariate delta method (e.g., Raykov & Marcoulides, 2004) as implemented in Mplus (Marsh, Lüdtke, et al. 2009; also see Mplus syntax in Supplemental Materials, Appendix 4). This opertionalization of contextual effects is potentially confusing in that it is not the effect of the L2 construct per se (e.g., class-average ability in the present investigation), but the difference between the L2 and L1 effects that constitutes the contextual effect.

Climate effects. Climate effects in the present investigation are based on reflective aggregations of L1 measures. For climate ratings all students within the same class are asked to rate a common construct (the climate of the classroom as a whole) rather than a characteristic that is idiosyncratic to each individual student (e.g., achievement, as in contextual effects). In this case, the climate effect is the effect of the aggregated L2 construct, but is not adjusted for corresponding L1 measure. The L1 ratings of climate are important in terms of estimating agreement among students within the class; if there is no significant agreement among students then it can be argued that the L2 aggregate does not reflect classroom climate. In contrast to contextual models, in climate models the L1 climate ratings have no substantive meaning in themselves with regards to climate effects. As in traditional multilevel models, climate effects are the effect of the aggregated L2 construct, not the difference between the corresponding L1 and L2 constructs. In the present investigation, the L1 climate ratings are modeled as part of the measurement component of the model, but not the structural component. Although the effects of the L1 measures may or may not be interpretable, or even relevant to other aspects of a particular study, they do not reflect climate effects. We leave as an open question whether the L1 climate ratings have any substantive role in the interpretation of the results, but caution that researchers who make substantive interpretations of L1 climate ratings must provide a theoretical and statistical rationale for doing so. However, it is critical that climate ratings are based on appropriately defined L2 constructs and not the L1 ratings. Indeed, the finding that the L1 climate ratings are nearly uncorrelated with other constructs is consistent with this rationale. In summary, it is critical that climate ratings are based on appropriately defined L2 constructs and not the L1 ratings.

Standardization and Effect Sizes. Mplus currently achieves standardisation in doubly-latent model separately for each level - treating them almost as multiple (separate) groups. This is reasonable when the researcher wants to evaluate these coefficients separately at L1 and L2. However for contextual studies, researchers need to consider coefficients between the two levels, so that the default standardized coefficients are not particularly useful. Following Marsh, Lüdtke, et al. (2009) we first standardized all L2 effects in relation to the total (L1 and L2) variance. Although we did this for both climate and contextual effects, it is important to reiterate that the contextual effects (because of the implicit group mean centering) are the difference between L1 and L2 effects of contextual variables, whilst climate effects are merely the effect of the L2 climate variable (see earlier discussion). Then we computed two measures of effect sizes. The first is a widely used measure proposed by Harker and Tymms (2004) for continuous level-2 predictors in MLMs, which they suggest is comparable with Cohen's *d* (Cohen, 1988):

$$ES1 = (2 * B * SD_{predictor})/\sigma_e \quad (1)$$

where B is the unstandardized regression coefficient in the MLM, $SD_{predictor}$ is the standard deviation of the predictor variable at L2, and σ_e is the residual standard deviation at L1. The resulting effect size describes the difference in the dependent variable between two L2 groups that differ by two standard deviations on the predictor variable. Alternatively, using the same notation (equation 1), Marsh, Lüdtke, et al. (2009) suggested it may be more appropriate to operationalize the effect size in relation to the total variance of the L1 variable rather than its residual (see Appendix 4 for Mplus syntax). In the published version of this study we only present ES2 (which we refer to as the ES).

Reliability of class-average responses. One problematic aspect of the manifest contextual analysis model is that the observed classroom-average $\overline{X}_{\bullet\bullet j}$ might be a highly unreliable measure of the unobserved classroom average because only small numbers of L1 students are sampled from each L2 classroom (O'Brien, 1990). Lüdtke, Marsh, et al. (2008; 2011; Marsh, Lüdtke, et al., 2009) introduced a multilevel latent covariate approach that takes into account sampling error when estimating group effects (see also Croon & van Veldhoven, 2007). In this approach the true group mean is considered as an unobserved latent

variable U_{xj} that is measured with a certain amount of precision by the group mean of the observed data (Asparouhov & Muthén, 2007). The precision is given by the intraclass correlation coefficient $\frac{\tau_x^2}{\tau_x^2 + (\sigma_x^2 / n_j)}$ where τ_x^2 is the variance between groups and σ_x^2 is the variance within groups. In the literature on reliability of multilevel data (Bliese, 2000) this measure is also sometimes called the ICC(2) and is used to determine the reliability of aggregated individual level data (e.g., the observed classroom average $\overline{X}_{\bullet\bullet j}$) in terms of sampling only a finite number of L1 units from each L2 unit. Thus, agreement among individuals within a group is assessed with the intraclass correlation coefficient (ICC1) and ICC2 as an estimate of the reliability of the group average (analogous to the reliability of a test score based on the Spearman-Brown equation):

$$\text{L2 Reliability } (\bar{X}_{\bullet j}) = \frac{n \cdot \text{ICC}}{1 + (n - 1) \cdot \text{ICC}}.$$

Thus, ICC2 can be interpreted as the reliability of the group mean in relation to sampling error. In most cases, the mean group size can be entered for n_j if not all groups are of the same size (see Searle, Casella, & McCulloch, 1992, on how to deal with pronounced differences in group size).

Grand mean and Within-Group Centering. In traditional multilevel contextual models based on manifest measures, researchers typically center L1 constructs in relation to the mean of the group to which the individual belongs ($\overline{X}_{\bullet ij}$ - $\overline{X}_{\bullet\bullet j}$) or the grand mean ($\overline{X}_{\bullet ij}$ - $\overline{X}_{\bullet\bullet\bullet}$; Enders & Tofighi, 2007; Kreft et al., 1995). In both approaches, the L1 effect is the same but the L2 effect is fundamentally different and the source of much confusion (see discussion by Lüdtke et al., 2009). Nevertheless, results based on one are a simple mathematical transformation of the other (see Raudenbush & Bryk, 2002).

For grand-mean centering , the L1 and L2 effects are correlated and so it is crucial to control for interindividual differences in L1 student responses when interpreting the contextual effect. Hence, the contextual effect is the partial L2 effect that controls for the effect of the L1 ratings. Thus the L2 effect in the grand-mean centered model is the expected difference between two students with the same L1 response in two different groups that differ by 1 L2 unit.

For group-mean centered variables, the L1 and L2 components are uncorrelated. Hence, the L2 effects of the aggregated L1 student responses are not controlled for interindividual differences in L1 responses. Because the L2 effect represents both L1 and L2 effects, the contextual effect is the difference between the L2 and L1 effects. Hence, unlike the grand-mean centering approach, the L2 effect is not a direct estimate of the contextual effect.

The choice of grand-mean and group-mean centering in contextual effects is typically arbitrary, so long as the interpretation is consistent with the choice of centering. However, for doubly latent MLSEMs the traditional decomposition of variables into within-group (L1) and between-group (L2) components results in an implicit within-group centering. Hence, a contextual effect in the doubly-latent MLSEM (with group mean centering) is present if the L2 regression coefficient is significantly different from the L1 regression coefficient. In the present investigation, a test of this hypothesis was accomplished by calculating an additional parameter—the difference between the corresponding L2 and L1 regression weights—that is a direct estimate of the contextual effect (see Mplus syntax in Appendix 4). Although this subtraction could easily be done by simple inspection, the important advantage of the approach used here is that it also provides a standard error of the estimate and facilitates the computation of standardized effect sizes.

Online Supplemental Materials

Appendix 3: Results of the Multilevel CFA and SEMs

All statistical analyses were done with Mplus (version 6; Muthén & Muthén, 2006-2010) using the doubly-latent MLSEM described in greater detail by Marsh, Lüdtke, et al. 2009; Lüdtke et al., 2008; 2011 (also see Appendix 2 for additional technical details and Appendix 4 for Mplus syntax). In order to facilitate interpretation of the coefficients in multilevel models and to reduce multicollinearity, all variables were standardized (M = 0, SD = 1). 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, 2006-2010) as well as the robust $\chi 2$ test statistic, inspection of parameter estimates, and the relative fit of different models in a nested or partially nested taxonomy of models designed a priori to evaluate particular aspects of interest rather than single models (Marsh, 2007a; Marsh, et al., 2009; also see Marsh, Hau, Balla & Grayson, 1998).

Supplemental Appendix 3a

Summary of Goodness of Fit Statistics 1

Model	CHI	df	CFI	TLI	RMSEA	SRMR	Description
Multilevel CFA Measurement Models: Tests of invariance							
MLCFAMM1	582	272	.982	.976	.023	.024 .102	No Invariance
MLCFAMM2	593	276	.981	.976	.023	.024 .102	L1 Invar
MLCFAMM3	607	282	.981	.976	.023	.024 .067	L1/L2 Invar
MLCFAMM4	616	286	.980	.976	.023	.024 .068	L1 Invar, L1/L2 Invar
Multilevel SEM Contextual Models[a]							
MLCFAMM4	616	286	.980	.976	.023	.024 .068	L1 Invar, L1/L2 Invar

Note. CHI= chi-square; df=degrees of freedom; CFI= Comparative fit index; TLI=Tucker-Lewis Index; RMSEA= Root Mean Square Error of Approximation. SRMR= Standardized Root Mean Square Residual; CFA=confirmatory factor analysis. L1 Invar = Invariance of self-concept factor loadings across time 1 and 2 at the individual student level. L1/L2 Invar = invariance of factor loadings across the individual and class levels for self-concept, social comparison focus and chaos climates.

[a]Because the SEM contextual model is a "full-forward" model, the goodness of fit, df, and number of estimated parameters are necessarily the same.

Supplemental Appendix 3B

Standardized Parameter estimates for Multilevel CFA Measurement Models (see Model MLCFAMM4 in Appendix 3A)

Within (W) Level—Individual Students

	SES		IQ		T1MAch_L1		T1MSC_L1		T2SocCmp_L1		T2Chaos_L1		T2MAch_L1		T2MSC_W	
	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE
Factor Loadings																
Indicator 1	.63	.04	.99	.00	.99	.00	.59	.01	.69	.02	.79	.02	.99	.00	.63	.02
Indicator 2	.55	.05					.75	.01	.73	.02	.84	.01			.79	.01
Indicator 3							.81	.01	.79	.02	.68	.02			.84	.01
Indicator 4							.57	.02	.71	.02	.66	.02			.64	.02
Indicator 5							.81	.01							.83	.01
Covariance Matrix																
SES	1.00															
IQ	.18	.03	1.00													
T1MAch_L1	.37	.04	.63	.02	1.00											
T1MSC_L1	.08	.04	.23	.03	.31	.03	1.00									
T2SocCmp_L1	-.07	.03	-.01	.04	-.01	.04	-.03	.03	1.00							
T2Chaos_L1	-.06	.05	.00	.04	.01	.04	-.08	.03	.01	.04	1.00					
T2MAch_L1	.27	.04	.56	.03	.65	.03	.35	.02	-.02	.03	-.04	.04	1.00			
T2MSC_L1	.10	.04	.19	.04	.27	.03	.65	.02	-.05	.03	-.17	.03	.39	.03	1.00	

Between (B) Level-Class

	Est	SE	Est	SE	Est	SE	Est	SE
	T2CHAS_L2		T2SocCmp_L2		T2MAch_L2		T2MSC_B	
Factor Loadings								
Indicator 1	.97	.02	.96	.03	.99	.00	.92	.14
Indicator 2	1.00	.00	.95	.03			1.00	.00
Indicator 3	.94	.03	1.00	.00			.90	.07
Indicator 4	.90	.04	.98	.03			.95	.13
Indicator 5							.90	.07
Covariance Matrix								
T2CHAS_L2	1.00							
T2SocCmp_L2	.42	.12	1.00					
T2MAch_L2	-.28	.13	.10	.11	1.00			
T2MSC_L2	-.04	.14	-.54	.13	-.39	.14	1.00	

Note. T1 = time 1 ; T2 = time 2 ; L1 = student level ; L2 = class level ; MAch = math achievement ; MSC = math self-concept ; SocCmp = Social-comparison classroom climate ; Chaos = chaos classroom climate ; est = parameter estimate ; SE = standard error. Values in parentheses are covariances or residual covariances, whist values not in parentheses are path coefficients (see Figure 1 for a representation of the multilevel structural equation model and Appendix 4 for Mplus syntax).

Supplemental Appendix 3C

Unstandardized Parameter estimates for Multilevel CFA Measurement Models (see Model MLCFAMM4 in Appendix 3A)

L1 -- Individual Student Level

	SES		IQ		T1MAch_L1		T1MSC_L1		T2SocCmp_L1		T2Chaos_L1		T1MAch_L1		T2MSC_L1	
	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE
Factor Landings																
Indicator 1	1.00	.00	1.00	.00	1.00	.00	1.00	.00	1.00	.00	1.00	.00	1.00	.00	1.00	.00
Indicator 2	.87	.11					1.27	.03	1.07	.04	1.07	.03			1.27	.03
Indicator 3							1.36	.04	1.14	.03	.88	.03			1.36	.04
Indicator 4							0.98	.04	1.04	.04	.84	.03			0.98	.04
Indicator 5							1.35	.04							1.35	.04
Covariance Matrix																
SES	.40	.07														
IQ	.21	.03	.99	.04												
T1MAch_L1	.23	.03	.63	.04	.99	.06										
T1MSC_L1	.03	.01	.14	.02	.18	.02	.35	.02								
T2SocCmp_L1	-.03	.02	-.01	.03	-.01	.03	-.01	.01	.42	.03						
T2CHAS_L1	-.03	.02	.00	.03	.01	.03	-.04	.01	.00	.02	.54	.03				
T2MAch_L1	.13	.02	.44	.03	.49	.04	.16	.01	-.01	.02	-.02	.02	.61	.05		
T2MSC_L1	.04	.02	.12	.02	.17	.02	.24	.02	-.02	.01	-.08	.01	.19	.02	.38	.02

L2 – Classroom Level

	T2SocCmp_L1		T2Chaos_L1		T1MAch_L1		T2MSC_W	
	Est	SE	Est	SE	Est	SE	Est	SE
Factor Loadings								
Indicator 1	1.00	.00	1.00	.00	1.00	.00	1.00	.00
Indicator 2	1.07	.04	1.07	.03			1.27	.03
Indicator 3	1.14	.03	.88	.03			1.36	.04
Indicator 4	1.04	.04	.84	.03			0.98	.04
Indicator 5							1.35	.04
Covariance Matrix								
T2SocCmp_L1	.10	.03						
T2CHAS_L1	.04	.01	.12	.02				
T2MAch_L1	.01	.01	-.04	.02	.15	.03		
T2MSC_L1	-.02	.01	-.00	.01	-.02	.01	.02	.01

Note. T1 = time 1 ; T2 = time 2 ; L1 = student level ; L2 = class level ; MAch = math achievement ; MSC = math self-concept ; SocCmp = Social-comparison classroom climate ; Chaos = chaos classroom climate ; est = parameter estimate ; SE = standard error. Values in parentheses are covariances or residual covariances, whilst values not in parentheses are path coefficients (see Figure 1 for a representation of the multilevel structural equation model and Appendix 4 for Mplus syntax). Factor loadings are presented in a compact form to conserve space; factor loadings for all indicators other than the ones designed to measure each factor are constrained to be zero.

Online Supplemental Materials

Appendix 4: Annotated Mplus Syntax files

```
TITLE:     New Appendix Model climate & Contexual Model--chaos & TSC SC & ACH;
DATA:        FILE IS TIMSSCLIM260710.dat;
VARIABLE:        NAMES ARE
     T2SocCmp6, T2SocCmp7, T2SocCmp8, T2SocCmp9
     T2Chaos1 T2Chaos2 T2Chaos3 T2Chaos4
     T1MSC1 ,T1MSC2, T1MSC3, T1MSC4, T1MSC5
     T2MSC1, T2MSC2 ,T2MSC3 ,T2MSC4, T2MSC5
     T1MAch        T2MAch
     SES1 SES2 INTEL;

cluster = id_CLASS ;


within = T1MSC1 T1MSC2 T1MSC3 T1MSC4 T1MSC5 T1MAch
   SES1 SES2 INTEL;
      !All variables that are strictly L1 need to be specified here. If there
      !were some strictly L2 variables, they would also need to be specified with a
      "between!="!statement.;

 Centering = grandmean(T1MSC1 T1MSC2 T1MSC3 T1MSC4 T1MSC5
   T1MAch SES1 SES2 INTEL);
         ! Only the variables that are strictly L1 need to be centered, given that all
         ! variables that are both L1 and L2 will be implicitly group mean-centered.

ANALYSIS: Type is twolevel ; estimator = MLR;
MODEL:

%within%
      ! within refers to the individual student level (L1);

      !achievement
T1MAch_W by T1MAch; T1MAch@.00;

      !self-concept T1
      !parameter estimates followed by the same number in ()
      !are constrained to be equal. Math self-concept factor
      !loadings are constrained to be the same over time and over levels;
   T1MSC_W BY T1MSC1@1 (1);
   T1MSC_W BY T1MSC2 (2);
   T1MSC_W BY T1MSC3 (3);
   T1MSC_W BY T1MSC4 (4);
   T1MSC_W BY T1MSC5 (5);

      !teacher Social Comparison Feedback;
   T2SocCmp_W  by T2SocCmp6@1 (26);
   T2SocCmp_W  by T2SocCmp7 (27);
   T2SocCmp_W  by T2SocCmp8 (28);
   T2SocCmp_W  by T2SocCmp9 (29);
      !chaos
   T2CHOS_W by T2Chaos1@1 (16);
   T2CHOS_W by T2Chaos2 (17);
   T2CHOS_W by T2Chaos3 (18);
   T2CHOS_W by T2Chaos4 (19);

      !achievement
   T2MAch_w by T2MAch; T2MAch@.00;
       T2MAch_w (Rvar_ACw);

         !self-concept T2
   T2MSC_W BY T2MSC1@1 (1);
   T2MSC_W BY T2MSC2 (2);
   T2MSC_W BY T2MSC3 (3);
   T2MSC_W BY T2MSC4 (4);
```

```
   T2MSC_W BY T2MSC5 (5);
      !correlated uniquenesses for same items at T1 & T2;
   T2MSC1 WITH T1MSC1;
   T2MSC2 WITH T1MSC2;
   T2MSC3 WITH T1MSC3;
   T2MSC4 WITH T1MSC4;
   T2MSC5 WITH T1MSC5;

      !Residual variance term used to compute effect sizes;
   T2MSC_W (Rvar_SCw);

      ! Reciprocal Effects Between MSC and MAch at student level;
   T1MSC_W on T1MAch_W;
   T2MAch_w on T1MSC_W;
   T2MAch_w on T1MAch_W;
   T2MSC_w on T1MSC_W ;
   T2MSC_W on T1MAch_w ;
   T2MSC_w on T2MAch_w  (b_WSCAC);

      !L1 climate variable correlated with other L1 constructs;
   T2MSC_w  with T2CHOS_W T2SocCmp_W ;
   T2MAch_w with T2CHOS_W T2SocCmp_W ;
   T1MSC_W  with T2CHOS_W T2SocCmp_W ;
   T1MAch_W with T2CHOS_W T2SocCmp_W ;
   T2SocCmp_W  WITH T2CHOS_W;

      ! SES Based on two indicators and then control for SES;
   SES BY SES2@1;
   SES BY SES1*0.866;
   T2MSC_w on SES;
   T2MAch_w on SES;
   T1MSC_w on SES;
   T1MAch_w on SES;
   T2SocCmp_W  on SES;
   T2CHOS_W on SES;

      !treat IQ as L1 single-item factor & and control for it;
   iq BY INTEL; INTEL@0;
   T2MSC_w on IQ;
   T2MAch_w on IQ;
   T1MSC_w on IQ;
   T1MAch_w on IQ;
   T2SocCmp_W  on IQ;
   T2CHOS_W on IQ;

%between%
      ! between refers to the individual student level (L1);

      !teacher social comparison reference (invariant over level);
   T2SocCmp_B by T2SocCmp6@1 (26);
   T2SocCmp_B by T2SocCmp7 (27);
   T2SocCmp_B by T2SocCmp8 (28);
   T2SocCmp_B by T2SocCmp9 (29);

      !classroom Chaos climate (invariant over level);
   T2CHOS_B by T2Chaos1@1 (16);
   T2CHOS_B by T2Chaos2 (17);
   T2CHOS_B by T2Chaos3 (18);
   T2CHOS_B by T2Chaos4 (19);

      ! Define T2MSC based on 5 indicators;
      !Self-concept T2
   T2MSC_B BY T2MSC1@1 (1);
   T2MSC_B BY T2MSC2 (2);
   T2MSC_B BY T2MSC3 (3);
   T2MSC_B BY T2MSC4 (4);
```

```
   T2MSC_B BY T2MSC5 (5);

      !constrain L2 residuals to be non-zero;
   T2SocCmp8@0;T2Chaos2@0;T2MSC2@0;

      !achievement T2
     T2MAch_B by T2MAch; T2MAch@.00;
    T2MSC_B on T2MAch_B (b_BSCAC);

   T2SocCmp_B WITH T2CHOS_B;
   T2MAch@0;
   T2MAch_B ON T2CHOS_B (b_ACCHA) ;
   T2MAch_B ON T2SocCmp_B (b_ACSoccmp) ;
   T2MSC_B  ON T2CHOS_B (b_SCCHA)  ;
   T2MSC_B  ON T2SocCmp_B (b_SCTSC)  ;
   T2CHOS_B (RVR_BCHA);
   T2SocCmp_B (RVR_BTSC);
   T2MAch_B (rvr_BT2A);

      !Residual Variances ;
   T2SocCmp6        * .01 (U1);
   T2SocCmp7        * .01 (U2);
   T2SocCmp8        * .01 (U3);
   T2SocCmp9        * .01 (U4);
   T2Chaos1       * .01 (U5);
   T2Chaos2       * .01 (U6);
   T2Chaos3       * .01 (U7);
   T2Chaos4       * .01 (U8);
   T2MSC1         * .01 (U9);
   T2MSC2         * .01 (U10);
   T2MSC3         * .01 (U11);
   T2MSC4         * .01 (U12);
   T2MSC5         * .01 (U13);
   T2MSC_B        * .01 (U14);

Model Constraint:
      !Residual Variances constrained to be non-zero;
U1 > 0;U2 > 0;U3 > 0;U4 > 0;U5 > 0;U6 > 0;U7 > 0;U8 > 0;U9 > 0;
U10 > 0;U11 > 0;U12 > 0;U13 > 0;U14 > 0;

      ! Variance estimates for self-concept and achievement used to
      ! computing Effect sizes; estimates were taken from Preliminary Analyses;
   new (VarSC_B); VarSC_B = .018;
   new (VarSC_W); VarSC_W = .394;
   new (VarAC_B); VarAC_B = .528;
   new (VarAC_W); VarAC_W = .473;

      !effect Chaos on SC
   new(stSCCHA);
   stSCCHA = b_SCCHA*(sqrt(RVR_BCHA)/sqrt(VarSC_W + VarSC_B));
   new(ES1SCCHA);
   ES1SCCHA = b_SCCHA*(2*sqrt(RVR_BCHA)/sqrt(Rvar_SCw));
   new(ES2SCCHA);
   ES2SCCHA = b_SCCHA*(2*sqrt(RVR_BCHA)/sqrt(VarSC_W));

      !effect Chaos on Ach
   new(stACCHA);
   stACCHA = b_ACCHA*(sqrt(RVR_BCHA)/sqrt(VarAC_W + VarAC_B));
   new(ES1ACCHA);
   ES1ACCHA = b_ACCHA*(2*sqrt(RVR_BCHA)/sqrt(Rvar_ACw));
   new(ES2ACCHA);
   ES2ACCHA = b_ACCHA*(2*sqrt(RVR_BCHA)/sqrt(VarAC_W));

      !effect TSC on SC
   new(stSCTSC);
   stSCTSC = b_SCTSC*(sqrt(RVR_BTSC)/sqrt(VarSC_W + VarSC_B));
```

```
   new(ES1SCTSC);
   ES1SCTSC = b_SCTSC*(2*sqrt(RVR_BTSC)/sqrt(Rvar_SCw));
   new(ES2SCTSC);
   ES2SCTSC = b_SCTSC*(2*sqrt(RVR_BTSC)/sqrt(VarSC_W));

      !effect SocCmp on Ach
   new(stACTSC);
   stACTSC = b_ACSocCmp*(sqrt(RVR_BTSC)/sqrt(VarAC_W + VarAC_B));
   new(ES1ACTSC);
   ES1ACTSC = b_ACSocCmp*(2*sqrt(RVR_BTSC)/sqrt(Rvar_ACw));
   new(ES2ACTSC);
   ES2ACTSC = b_ACSocCmp*(2*sqrt(RVR_BTSC)/sqrt(VarAC_W));

      !NEW equations FOR bflpe-AC;
       new(bflpeAC);
       bflpeAC = b_BSCAC - b_WSCAC;
         new(stbfAC);
      stbfAC = bflpeAC*(sqrt(rvr_BT2A)
       /sqrt(VARSC_W +VARSC_B));    ! varT2MSC_W + varT2MSC_W);
       new(ES1BFLPE);
   ES1BFLPE= bflpeAC*(2*sqrt(rvr_BT2A) /sqrt(Rvar_SCw));

       new(ES2BFLPE);
   ES2BFLPE= bflpeAC*(2*sqrt(rvr_BT2A) /sqrt(VARSC_W));

OUTPUT: sampstat tech1;
```