Contemporary Educational Psychology 34 (2009) 120–131

Contents lists available at ScienceDirect

Contemporary Educational Psychology

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

Theoretical Analysis

Assessing the impact of learning environments: How to use student ratings of classroom or school characteristics in multilevel modeling

Oliver Lüdtke [a,*], Alexander Robitzsch [b], Ulrich Trautwein [a], Mareike Kunter [a]

[a] Max Planck Institute for Human Development, Center for Educational Research, Lentzeallee 94, 14195 Berlin, Germany
[b] Institute for Educational Progress, Berlin, Germany

ARTICLE INFO

Article history:
Available online 22 January 2009

Keywords:
Learning environment
Multilevel analysis
Agreement
Climate
Student ratings

ABSTRACT

In educational research, characteristics of the learning environment (e.g., social climate, instructional quality, goal orientation) are often assessed via student reports, and their associations with outcome variables such as school achievement or student motivation then tested. However, studying the effects of the learning environment presents a series of methodological challenges. This article discusses three crucial elements in research that uses student reports to gauge the impact of the learning environment on student outcomes. First, from a conceptual point of view, it is argued that ratings aggregated at the relevant level (e.g., class or school level), and not individual student ratings, are of primary interest in these studies. Second, the reliability of aggregated student ratings must be routinely assessed before these perceptions are related to outcome variables. Third, researchers conducting multilevel analyses need to make very clear which centering option was chosen for the predictor variables. This article shows that conclusions about the impact of learning environments can be substantially affected by the choice of a specific centering option for the individual student ratings.

1. Introduction

A key assumption of most research in the educational context is that cognitive, motivational, emotional, and behavioral student outcomes are substantially shaped by features of the learning environment, such as learning climate (Fraser & Fisher, 1982), instructional quality (Greenwald, 1997), and classroom goal structures (Ames, 1992; Karabenick, 2004). Empirical studies assessing these characteristics of the learning environment typically draw on one or more of three data sources (Anderson, 1982; Fraser, 1991; Turner & Meyer, 2000): external observers, teachers, or students.[1]

Each perspective has specific methodological and theoretical advantages and disadvantages. In general, it is argued that observer ratings provide the most objective accounts of specific classroom features because the raters are not directly involved in the teaching process (e.g., Walberg & Hartel, 1980). However, this approach is also the most cost- and labor-intensive, and not all constructs are easily observable to outsiders (e.g., goal structures). Teachers, with their professional training and knowledge, are experts on various instructional approaches. Because they are responsible for guiding the instructional process, moreover, they would seem to be the ideal source for reports on instructional practices (Mayer, 1999; Porter, 2002). However, the validity of teacher self-descriptions may be compromised and biased by teaching ideals or self-serving strategies (Wubbels, Brekelmans, & Hooymayers, 1992). Students are exposed to a variety of teachers in different subjects over their school careers, and may therefore be considered experts on different modes of teaching (Clausen, 2002; De Jong & Westerhof, 2001). From a phenomenological point of view, moreover, students' ratings are the most appropriate source of data for assessing the learning environment: a given student's behavior can be assumed to be more affected by his or her interpretation of the classroom context than by any objective indicator of that context. At the same time, given the idiosyncratic nature of students' perceptions of their learning environment, the reliability of students as data sources has repeatedly been questioned (Aleamoni, 1999; Marsh & Roche, 1997).

Because of the conceptual advantages of student reports, and because student reports are often more easily obtained than reports from teachers or external observers, the use of student reports to assess characteristics of the learning environment has recently flourished (e.g., Frenzel, Pekrun, & Goetz, 2007; Friedel, Cortina, Turner, & Midgley, 2007; Kunter, Baumert, & Köller, 2007). However, there are serious conceptual and methodological challenges that need to be addressed before student ratings can properly be used to gauge the effects of characteristics of the learning environment. These challenges have not yet received sufficient

* Corresponding author.
E-mail address: luedtke@mpib-berlin.mpg.de (O. Lüdtke).

[1] In the following, the term "student" refers to learners who rate different aspects of their learning environment (e.g., in schools, colleges, etc.).

0361-476X/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.cedpsych.2008.12.001

research attention. In this article, we discuss three major methodological challenges facing researchers using student ratings and describe how they can be addressed. Please note that this article is not intended to provide a general introduction to multilevel modeling. Readers who would like to learn more about multilevel modeling in general are referred to the excellent introductory books on this topic (in increasing order of technical sophistication: Bickel, 2007; Hox, 2002; Kreft & de Leeuw, 1998; Raudenbush & Bryk, 2002; Snijders & Bosker, 1999).

2. Choosing the appropriate level of analysis in learning environments research

Student questionnaires have been used to collect information about learning environments in a number of research areas. In the following, we use the term "learning environments" to describe institutionalized and naturally occurring group settings that stimulate learning in students (e.g., schools, classes, small groups). In research on technology-based learning environments, the term is used with a different connotation, to mean technological devices thought to facilitate learning in students (de Corte, 2001).

The first studies drawing on student perceptions to assess aspects of the learning environment were conducted in the field of climate research (Anderson, 1982; Moos, 1979; see Fraser, 1991, for an overview), and administered newly developed instruments such as the Learning Environment Inventory (Anderson, 1973), the Classroom Environment Scale (Moos & Trickett, 1974), and the Individualized Classroom Environment Questionnaire (Fraser, 1980). These instruments were the result of factor analytic studies in which students rated their learning environment on several dimensions, such as student/teacher relationships (e.g., teacher support, involvement in discussions, personalization), personal development (e.g., speed, difficulty, competitiveness), and system maintenance and change (e.g., rule clarity, teacher control, differentiation). Typical items in these instruments are: "The teacher takes a personal interest in the students" (teacher support), "The pace of the class is rushed" (speed), and "There is a clear set of rules for students to follow" (rule clarity). The main objective was to investigate the relationship between these different climate dimensions and student outcomes such as achievement and attitudes (e.g., Fraser & Fisher, 1982).

A second strand of research that typically employs student ratings to assess learning environments is research on teaching effectiveness (for an overview, see Seidel & Shavelson, 2007). Several such studies have been conducted in university settings, with student ratings being used to evaluate the outcomes of certain courses or instructors (e.g., Aleamoni, 1999; Marsh & Roche, 1997). One prominent instrument for assessing teaching effectiveness is the Students' Evaluation of Educational Quality (SEEQ; Marsh & Bailey, 1993), which captures dimensions of instruction such as organization (e.g., "The course materials were well-prepared and carefully explained") and enthusiasm (e.g., "The instructor was dynamic and energetic in conducting the course"). Furthermore, in large-scale educational assessments evaluating whole schools, systems, or countries, student questionnaires tapping school or class variables are used to gain insights into mediating processes that may account for differences in educational effectiveness (Anderson, Ryan, & Shapiro, 1989; OECD., 2001).

A third important research area drawing on student perceptions of the learning environment concerns students' motivational development. Research on achievement goal theory has proved particularly fruitful in recent years. Achievement goal theory posits that individuals have different purposes for engaging (or not engaging) in activities, and that these purposes, or goal orientations, are systematically linked to behavioral differences (Ames, 1992). One central hypothesis is that teacher endorsement of different types of goals (mastery versus performance goals) affects students' patterns of learning and behavior (e.g., Karabenick, 2004). Recently, the focus of this research has shifted to the "classroom goal structure," which is assumed to affect not only individual goal orientations, but also other student characteristics (see Miller & Murdock, 2007). Most studies taking this approach have used the Pattern of Adaptive Learning Scales instrument (PALS; Midgley et al., 2000), which assesses classroom mastery goal structure (e.g., "In our class, how much you improve is really important"), classroom performance-approach goal structure (e.g., "In our class, getting good grades is the main goal"), and performance-approach goal structure (e.g., "In our class, showing others that you are not bad at class work is really important").

In all of these areas, the level of analysis (Raudenbush & Bryk, 2002) is crucial. Students who report on their environment are nested within classes or courses, resulting in two different levels that might be considered. The first level is that of individual students (also called "student level" or "level-1"); the second level is that of classes or schools (often referred to as "class level" or "level-2"). When researchers obtain data from many students within a class or school, they can aggregate these ratings at the class/school level to yield a measure of the "shared perception of the environment" (i.e., the mean ratings of students in each class/school). They need to decide on the appropriate procedure for dealing with these hierarchical data from a conceptual perspective. For instance, disregarding complications such as causal direction, if a researcher is interested in the effects of a supportive class climate on student motivation, is it appropriate to examine the relationship between a student's individual perception of classroom climate and his or her motivation? Or would it make more sense to aggregate student perceptions at the classroom or school level and to analyze the association of the aggregated score with the outcome variables?

Both individual- and class-level student ratings have been considered in educational research. Many early analyses of the relationship between learning climate and school achievement focused exclusively on the individual level (e.g., Walberg, 1972). Likewise, much research on achievement goals has failed to specify which level of analysis is at the center of interest. In their overview of recent studies on goal structure effects, Miller and Murdock (2007) and Miller (2006) found that 16 of the 31 field studies included in their review correlated individual students' perceptions of the classroom goal structure (mastery versus performance goal structure) with one or more motivational outcome variables, but did not include any class-level analyses. Miller and Murdock warned that this analytical strategy may not be the ideal design for investigating effects of classroom environments (see also Lau & Nie, 2008).

Clearly, the analytical level appropriate in any given case depends on the research question addressed. The crucial question is whether researchers are interested in differences between students or differences between learning environments. Murray's (1938) need–press model provides an important theoretical basis for the differentiation between individual and shared perceptions. Murray assumed that behavior results from the interplay of environmental influences, or "press," and individual needs. Environmental influences are factors that may promote or impede the satisfaction of individual needs. In describing aspects of the environment, Murray distinguished between alpha press, "the press that actually exists as far as scientific discovery can determine it," and beta press, "the subject's own interpretation of the phenomena that is perceived." Stern (1970) enhanced that distinction, differentiating between truly idiosyncratic private beta press, perceptions that individual members of a group have about an environment, and mutually shared consensual beta press, perceptions that members of a group shared about an environment. Transfer-

ring this approach to the school context, it is possible to distinguish the individual student perspective and the collective perspective of the class as a whole. Student ratings thus allow aspects of the learning environment to be conceptualized in two different ways (Cronbach, 1976). Adopting a multilevel perspective, one can differentiate between the individual level, in which student ratings represent an individual student's perception of the learning environment (private beta press), and the classroom level, reflecting the average perception of the learning environment in a given class (consensual beta press).

What, then, are educational researchers interested in? Researchers who focus on interindividual differences among students are often interested in "private beta press" – for instance, whether individual students' perceptions of their classroom environment (e.g., whether or not they feel supported by the teacher or well integrated in the class) are related to different cognitive, motivational, and behavioral outcomes. Individual perceptions of the classroom are thus central here. In contrast, research on the effects of the learning climate typically attempts to describe how different environments affect the outcomes of students in different classes, and research on teaching effectiveness investigates the effects of differences between teachers/schools. In this context, the main purpose of measuring individual students' perceptions of their learning environments is to assess a group-level or class-level construct (e.g., the class climate or the teaching quality of a particular teacher as perceived by his/her students). Hence, the primary unit of analysis should be the group (e.g., class or school) and not the individual. In his seminal paper on multilevel issues in educational research, Cronbach (1976) was very clear about the role of students' perceptions when assessing aspects of the learning environment. In a discussion of the Learning Environment Inventory (LEI; Anderson, 1973) he argued that: "The purpose of the LEI 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. 9.18).

From this perspective, the main purpose of collecting individual students' ratings of their class or school is to assess aspects of environments that are clearly located at the group level (e.g., class or school level). Thus, students are regarded as informants on their learning environment, in the sense of multiple observers providing data on one construct. At the individual level, the measurements refer to the phenomenology of the student. At the class level, however, they refer to differences between classes. If educational researchers want to gain insights into the effects of learning environments, they have no choice but to use aggregated student data, or other original class-level measures, such as teacher reports or data from independent observers.

As a consequence, educational researchers interested in the effects of different aspects of students' learning environments need to observe a large enough number of learning environments (e.g., classes or schools) that exhibit sufficiently large differences in the characteristics examined. The number of classes or schools needed for such a study depends on various aspects (e.g., number of students per environment, variance in predictor and outcomes variables), but for most research questions a sizeable number of classes/schools (at least 40–50 groups) is typically needed to obtain stable estimates of group effects (Hox, 2002; see Spybook, 2008, for an overview of sample size requirements). Individual, idiosyncratic perceptions do not allow any conclusions to be drawn about the effects of environments. Studying the effects of individual student ratings within classrooms may be interesting for many research questions. However, researchers need to be aware that, in this case, they are not dealing with differences between environments, but merely with idiosyncratic student perceptions.

3. Psychometric properties of aggregated student ratings

Once researchers have identified the classroom or school level as the theoretically appropriate level, they need to investigate the psychometric properties of the aggregated ratings. In other words, having theoretically determined that their construct is a class-level or school-level construct, researchers need to show that the aggregated student responses provide a psychometrically sound measure of that construct. In the same way as indices such as Cronbach's alpha are used to describe the reliability of multi-item scales, researchers should report the psychometric quality of their aggregated constructs. In short, before aggregating student perceptions of learning characteristics at the class or school level, researchers must determine whether it really makes sense to form an aggregate variable. This will only be the case if the variation of student responses within or across classes/schools is in some way systematic.

Although there is a long tradition within educational research (see Kane & Brennan, 1977; Kane, Gillmore, & Crooks, 1976, for studies on the reliability of class means) of studying the reliability of group means in the framework of generalizability theory (Brennan, 2001), these ideas have not found their way into current research using student ratings of learning environments. As a consequence, reliabilities for aggregated student ratings are very rarely reported (see also Miller & Murdock, 2007). A simple way of determining whether aggregated individual-level ratings are reliable indicators of group-level constructs is to use the intraclass correlations $ICC(1)$ and $ICC(2)$ (Bliese, 2000; Raudenbush & Bryk, 2002). These indices are based on a one-way analysis of variance with random effects, where the individual-level rating at level-1 is the dependent variable and the grouping variable (e.g., class, school) is the independent variable. The $ICC(1)$ is defined as follows:

$$ICC(1) = \frac{MS_B - MS_W}{MS_B + (k-1)MS_W} \quad (1)$$

where MS_B is the between-group mean square, MS_W is the within-group mean square, and k is the group size. The $ICC(1)$ indicates the proportion of total variance that can be attributed to between-class differences. In the case of student ratings, the $ICC(1)$ can be seen as a measure of effect size indicating the extent to which individual ratings are affected by the learning environment (e.g., the degree to which students' ratings of classroom features are affected by the fact that they are in different classes). If there is no systematic variation between classes, it makes no sense to investigate effects at class level. Researchers typically use standard software such as SPSS or a multilevel software package to estimate the $ICC(1)$.[2] Whereas the $ICC(1)$ refers to individual students' ratings, the $ICC(2)$ provides an estimate of the reliability of class-mean ratings. The $ICC(2)$ is estimated by applying the Spearman-Brown formula to the $ICC(1)$, with k being the number of students per class (Bliese, 2000; Snijders & Bosker, 1999)

$$ICC(2) = \frac{k \times ICC(1)}{1 + (k-1) \times ICC(1)} \quad (2)$$

[2] In the literature on multilevel modeling, the $ICC(1)$ is also frequently defined as $ICC = \frac{\tau^2}{\tau^2+\sigma^2}$, where τ^2 is the variance between classes and σ^2 the variance within classes. The variance components τ^2 and σ^2 are then estimated directly using maximum-likelihood procedures. Given that $E(MS_W) = \sigma^2$ and $E(MS_B) = \sigma^2 + k\tau^2$, the results are very similar when the sample is large and the data are balanced (Snijders & Bosker, 1999).

As is apparent from Eq. (2), the reliability of the class-mean rating as estimated by the *ICC*(2) increases with the number of students *k*. In other words, the more students in a class provide ratings, the more accurately the class-mean rating will reflect the true value of the construct being measured. If not all classes are of the same size, the mean class size can generally be entered for k (see Bliese, 2000, on how to deal with pronounced differences in class size). The *ICC*(2) has to be calculated manually because it is not included in most software packages. Depending on the measures used in the multilevel analysis, values between .70 and .85 are usually taken to indicate acceptable levels of reliability (LeBreton & Senter, 2008; Lüdtke, Trautwein, Kunter, & Baumert, 2006). Fig. 1 illustrates the calculation of the *ICC*(1) and *ICC*(2) using a small hypothetical data set in which students' ratings are nested within classes. In this hypothetical example, 10 students judge an aspect of their classroom environment on an item with a 5-point response format. In educational research, for example, 10 students in each class might be asked to rate the quality of the homework assigned by their teacher. In this example, it is obvious that the mean ratings for homework quality differ systematically between classes, ranging from 2.3. to 3.7. This systematic between-class variation is reflected in the *ICC*(1) of .43, which indicates that 43% of the total variation found in all student ratings can be attributed to the fact that students are nested in certain classes. Combining the *ICC*(1) of .43 with an average number of 10 students providing ratings per class yields an *ICC*(2) of .80, indicating a sufficient degree of reliability of the class-mean ratings.

It is worth noting that Eq. (2), which determines the reliability of the observed group mean, says nothing about the reliability of the level-1 measure. In general, measurement error at level-1 results in lower reliability of the group means (Kamata, Bauer, & Miyazaki, 2008; Raudenbush & Bryk, 2002). However, Eq. (2) does not differentiate between level-1 variance that is due to measurement error and level-1 variance that is due to true differences between individuals.

4. Analysis of learning environments using multilevel modeling

Having established the reliability of aggregated perceptions, researchers can start to investigate the effects of the learning environment on outcome variables. The method most suitable for these analyses is multilevel modeling (MLM; also known as hierarchical linear modeling, HLM), a general form of regression analysis that provides a powerful methodology for handling the hierarchical data that are typical of research on learning environments (see Raudenbush & Bryk, 2002; Snijders & Bosker, 1999). However, multilevel modeling provides researchers with several modeling options that—if used inappropriately—may compromise the validity of the results. In the context of aggregated student ratings, two main decisions must be made. The first concerns the appropriate level of analysis. MLM allows the effects of individual students' perceptions and class-average perceptions of the learning environment to be modeled at the individual level, at the class level, or simultaneously at both levels (e.g., Karabenick, 2004; Trautwein, Lüdtke, Schnyder, & Niggli, 2006; Urdan, Midgley, & Anderman, 1998). As described above, it is imperative for researchers examining the effects of learning environments to focus on the class level by introducing class-level predictor variables (e.g., aggregated student ratings of homework quality). However, it is also possible (and frequently seen in the literature) for researchers to simultaneously introduce individual-level predictor variables to examine research questions addressing interindividual differences at the individual level (e.g., individual student ratings of homework quality).

When researchers decide to include student ratings of their learning environment at the individual level, they face a second crucial decision: What is the appropriate centering option for students' ratings at the individual level? In most MLM models, level-1 predictors are not used in their raw metric, but are subjected to different transformations to make the interpretation of the coefficients and variance components more meaningful (see Raudenbush & Bryk, 2002). Basically, two centering options can be distinguished (see Fig. 2 for an illustration). First, student ratings of a given feature can be adjusted to the mean ratings of that feature in the cluster to which the student belongs (centering at the group mean or centering within cluster; CWC). As shown in the fifth column of Fig. 2, the respective class mean is subtracted from the individual ratings (raw data in second column). Within-class and between-class effects of students' perceptions of the learning environment on student outcomes can thus be disentangled. Most importantly, when CWC is applied, the between-class effects of the aggregated student perceptions are not controlled for interindividual differences in student perceptions of the learning environment. Second, student ratings of a given feature can be adjusted to the mean ratings of that feature in the whole sample (centering at the grand mean; CGM). In other words, each student's relative position on the predictor variable within the whole sample is determined by subtracting the grand mean from each individual rating (see the sixth column of Fig. 2). When CGM is applied, it is crucial for interpreting the regression coefficients that interindividual differences in student perceptions of the learning environment are taken into account when the between-class effect of the

	Rating given by student										
	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	$\bar{X}_{\bullet j}$
Class 1	2	2	1	2	3	3	3	2	2	3	2.3
Class 2	3	2	3	2	3	3	2	3	4	3	2.8
Class 3	3	2	3	2	3	3	3	4	2	3	2.8
Class 4	2	3	4	2	3	2	4	3	2	4	2.9
Class 5	5	4	3	4	2	4	4	3	4	4	3.7

$$MS_W = \frac{\sum_{j=1}^{5}\sum_{i=1}^{10}(X_{ij} - \bar{X}_{\bullet j})^2}{5\times(10-1)} = .54 \qquad ICC(1) = \frac{MS_B - MS_W}{MS_B + 9\times MS_W} = \frac{2.55 - .54}{2.55 + 9\times .54} = .43$$

$$MS_B = \frac{\sum_{j=1}^{5}(\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet})^2}{5-1} = 2.55 \qquad ICC(2) = \frac{10\times ICC(1)}{1 + 9\times ICC(1)} = \frac{10\times .43}{1 + 9\times .43} = .80$$

Fig. 1. Illustration of the calculation of the *ICC*(1) and *ICC*(2) for five classes, each with 10 students (i.e., students nested within classes), rating a feature of the classroom environment on a 5-point scale.

	Student	Raw Data X_{ij}	Class Mean $\overline{X}_{\cdot j}$	Grand Mean $\overline{X}_{\cdot\cdot}$	CWC $X_{ij} - \overline{X}_{\cdot j}$	CGM $X_{ij} - \overline{X}_{\cdot\cdot}$
Class 1	1	2	2.3	2.9	2 – 2.3	2 – 2.9
	⋮	⋮	⋮	⋮	⋮	⋮
	10	3	2.3	2.9	3 – 2.3	3 – 2.9
Class 2	1	3	2.8	2.9	3 – 2.8	3 – 2.9
	⋮	⋮	⋮	⋮	⋮	⋮
	10	3	2.8	2.9	3 – 2.8	3 – 2.9
	⋮	⋮	⋮	⋮	⋮	⋮
Class 5	1	5	3.7	2.9	5 – 3.7	5 – 2.9
	⋮	⋮	⋮	⋮	⋮	⋮
	10	4	3.7	2.9	4 – 3.7	4 – 2.9

Fig. 2. Illustration of group-mean centering (CWC) and grand-mean centering (CGM) of individual student ratings using the data set from Fig. 1.

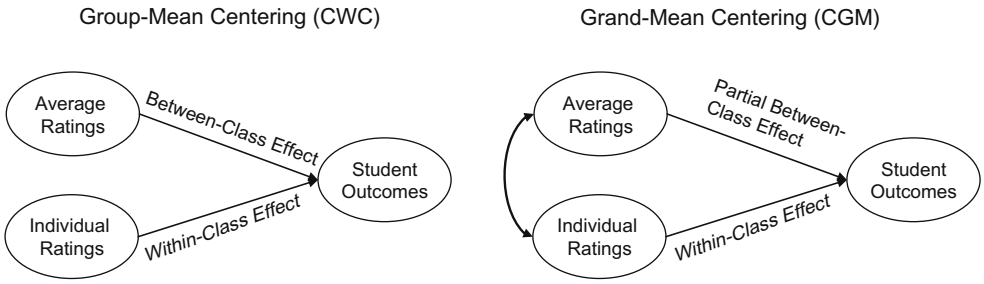

Fig. 3. Schematic comparison of the basic multilevel model for the analysis of students' ratings of the learning environment when individual ratings are centered at the group mean (CWC) or at the grand mean (CGM). Note that, in the grand-mean centered case, individual and aggregated perceptions are correlated, yielding a partial between-class coefficient.

aggregated student perceptions on student outcomes is estimated. Hence, the between-class effect can be regarded as a partial effect controlled for the effect of the individual ratings. Fig. 3 provides a schematic comparison of both centering options for the basic multilevel model used to analyze students' perceptions of their learning environment. As can be seen, the within-class effect remains the same in both models. However, the between-class effect in the CGM model is a partial regression coefficient, whereas the coefficient in the CWC model is not adjusted for differences in individual perceptions. This is also expressed in the fact that individual and average ratings are uncorrelated in the CWC model but correlated in the CGM model. It is important to add that under certain circumstances (when only the level of the outcome variable is allowed to vary; random-intercept models) the partial between-class effect of the CGM model can be obtained from the estimates of the CWC model and vice versa (e.g., Kreft, de Leeuw, & Aiken, 1995). The Appendix provides a detailed description of the mathematical properties of the basic multilevel model for analyzing the effects of student ratings and explains how the coefficients of the CWC model and the CGM model are related.

4.1. Lack of consistency in research practice

The literature on multilevel modeling (e.g., Enders & Tofighi, 2007; Raudenbush & Bryk, 2002; Snijders & Bosker, 1999) highlights the differences between the two centering options, but little guidance is offered on centering when student ratings of the learning environment are used simultaneously as predictor variables at the individual and class level. Not surprisingly, then, a review of studies using multilevel modeling to analyze the effect of student ratings of the learning environment reveals that there is no consensus among authors—or journal reviewers—on whether CGM or CWC should be used. Several researchers have centered individual students' ratings at their grand mean (e.g., Kunter et al., 2007; Lüdtke, Köller, Marsh, & Trautwein, 2005; Trautwein et al., 2006; Urdan, 2004; Wendorf & Alexander, 2005). Others have applied CWC to individual student ratings (e.g., Church, Elliot, & Gable, 2001; Kaplan, Gheen, & Midgley, 2002; Karabenick, 2004; Ryan, Gheen, & Midgley, 1998; Turner et al., 2002; Urdan et al., 1998).[3] In some studies, however, the authors did not indicate which centering option was used (e.g., Trouilloud, Sarrazin, Bressoux, & Bois, 2006; Wong, Young, & Fraser, 1997). Clearly, this inconsistency in centering is reason for concern: all of these studies were essentially interested in the same research question, namely the effect of features of the learning environment on student outcomes.

To illustrate the difference between CGM and CWC, we present two examples from recent research using student ratings to assess the effects of learning environments. Karabenick (2004) related college students' patterns of help-seeking behavior to individual and aggregated ratings of their classes' achievement goal structure. Karabenick decided to use CWC of level-1 predictors in his multilevel analyses because the "purpose was to differentiate between-class from within-class variation in perceived goal structure" (p. 575). In one multilevel model, students'

[3] Alternatively, researchers sometimes decide to include only the aggregated student perceptions in the MLM (e.g., Lau & Nie, 2008; Papaioannou, Marsh, & Theodorakis, 2004). This approach essentially corresponds to the CWC option, because the group means of the outcome variable are not adjusted for differences in the individual students' perceptions.

Table 1
Effects of students' ratings of the quality of their homework assignments on homework effort

Homework effort	Group-mean centered		Grand-mean centered	
	Coefficient	SE	Coefficient	SE
Fixed effect				
$\hat{\gamma}_{00}$ Intercept	0.05	0.03	0.05	0.03
$\hat{\gamma}_{01}$ Homework quality (average)	0.43	0.06	0.19	0.06
$\hat{\gamma}_{10}$ Homework quality (student)	0.24	0.02	0.24	0.02
$\hat{\gamma}_{20}$ Gender: male	−0.09	0.04	−0.09	0.04
$\hat{\gamma}_{30}$ Conscientiousness	0.43	0.02	0.43	0.02
$\hat{\gamma}_{40}$ Basic cognitive abilities	−0.01	0.02	−0.01	0.02
	Var. comp.			Var. comp.
Random effect				
$Var(u_{0j})$	0.03		0.03	
$Var(r_{ij})$	0.60		0.60	

Note: N(level-1) = 1501. N(level-2) = 93. Average cluster size = 16.14. *SE* = standard error. Var. comp. = variance component. All parameter estimates except the intercept and $\hat{\gamma}_{40}$ are statistically significantly different from zero ($p < .05$). In the group-mean centered model, individual students' ratings of homework quality were centered at their group mean; in the grand-mean centered model, they were centered at their grand mean.

help-seeking avoidance pattern was related to students' ratings of three dimensions of the class goal structure: mastery–approach, performance–approach, and performance–avoidance. Including these three dimensions as predictor variables in a multilevel model, Karabenick found a statistically significant positive effect (.773) of the aggregated student perceptions of performance–avoidance on students' help-seeking avoidance pattern. The corresponding effect of the individual ratings at the individual level was .230. Hence, if CGM rather than CWC had been used to center the individual ratings of the performance–avoidance goal structure in the classroom, the between-class coefficient (as a partial regression coefficient) would be .543, considerably lower than the effect reported for CWC, which does not control for individual differences in students' perceptions of the performance–avoidance goal structure.[4]

In another example, Wendorf and Alexander (2005) examined the effect of individual- and class-level ratings of fairness (procedural, interactional, and distributive fairness) on students' satisfaction with their instructor. In their multilevel analyses, all individual-level measures of fairness were grand-mean centered. Thus, the effect of the aggregated class-level perceptions on satisfaction was controlled for interindividual differences in individual students' ratings of fairness. Both the aggregated perceptions of interactional fairness (e.g., "This instructor has interacted with students appropriately and fairly") (.901) and the corresponding individual-level ratings (.510) had a statistically significantly positive effect on students' instructor satisfaction. If the authors had decided not to report the partial effect for the aggregated perceptions of interactional fairness, and had applied CWC to the individual perceptions of fairness (or had they not included the individual perceptions as level-1 predictor variables), the resulting between-class regression coefficient would have been 1.411, a considerably larger effect.

As shown by these two examples, dramatic differences can emerge between CGM and CWC. First, the size of the effect of the "classroom environment" is often substantially affected by the choice of a specific centering option. Second, although it was not the case in our two examples, the choice of a specific centering option may even affect the statistical significance of a regression coefficient. This is an unfortunate situation. From a mathematical perspective it makes no difference whether CGM or CWC is used—both the unadjusted and the adjusted between-class regression coefficient can be obtained from either model (see Eq. (A.7) in the Appendix). From the perspective of testing theoretical hypotheses (e.g., the effect of a certain classroom feature), however, the choice of either CGM or CWC is of central relevance. Is there a "correct" centering option in research on learning environments? As emphasized by Enders and Tofighi (2007), and several other methodologists (Hofmann & Gavin, 1998; Kreft et al., 1995; Snijders & Bosker, 1999), "the decision to use CGM or CWC cannot be based on statistical evidence, but depends heavily on one's substantive research questions" (p. 135). In fact, different strategies should be applied depending on the variables used and the research question addressed. We will come back to the issue of centering individual student ratings after presenting a data example from homework research that illustrates the issues raised thus far.

5. A data example: the effects of homework assignment

The present application is a reanalysis of a large-scale, multilevel homework study with 1501 eighth graders from 93 classes (Trautwein et al., 2006). The main goal of the study was to examine the effects of homework assignments (here: French as foreign language) on homework motivation and homework behavior. In this sample, all students in the same class were assigned the same homework. The authors argued that the quality of homework assignments affects student outcome variables. They obtained student ratings of homework quality (e.g., "French homework really makes us think") as well as student reports on homework motivation and effort (e.g., "I always try to finish my French homework"), and used multilevel modeling to predict homework motivation and effort. In the following, we use this example to illustrate our suggestions of how to relate student perceptions of the learning environment to student outcomes.

In the first step, one needs to consider the correct conceptualization of the variables. Is homework quality a student level or a class-level variable? From a theoretical point of view, Trautwein and colleagues (2006) were primarily concerned with the differing effects of homework quality as a feature of the learning environment across teachers/classes. Accordingly, and in line with the conceptual arguments presented in the first part of this article, homework quality should be treated as a class-level variable. Hence, although homework quality was assessed by student report, it was subsequently also aggregated at the class level and used as class-level variable. More specifically, in this design, (a) the rating of an individual student can be interpreted as that student's perception of the quality of the homework assigned by the teacher and (b) the class-average response of students can be interpreted as a collective perception of homework quality, in which individual idiosyncracies are averaged out.

The next step was to calculate the reliability of the aggregated student ratings. Can the students' ratings of homework quality indeed be used as a class-level variable? Applying the mixed procedure in SPSS 15 (Peugh & Enders, 2005), we estimated an unconditional multilevel model (i.e., a model in which only the individual student ratings of homework quality were included as the outcome variable and the class was introduced as a random factor) using restricted maximum likelihood. The unconditional model decomposes the total variance in student ratings into variance attributable to students (within-class variance) and variance

[4] Here, we used the relation shown in the Appendix that the partial between-group regression coefficient of the CGM model can be calculated by taking the difference between the between-group regression coefficient and the individual level coefficient of the CWC model (see Eq. (A.7)): .773−.230 = .543.

attributable to classes (between-class variance). This model showed that 17% of the total variance in students' perceptions of their homework quality was located between classes, indicating that the mean student ratings of homework quality varied across classes.[5] This translates directly into an *ICC*(1) of .17 for homework quality (see formula in Footnote 2). The *ICC*(1) rarely show values greater than .30 in educational and organizational research (Bliese, 2000; James, 1982). Whereas the *ICC*(1) indicates the reliability of an individual student's rating, the *ICC*(2) provides an estimate of the reliability of the class-mean rating. Applying the Spearman-Brown correction with an average number of ratings per class of 16.1 gives an *ICC*(2) of .77 for homework quality, indicating sufficient reliability of the aggregated ratings.

Having demonstrated that students' ratings of their homework quality can be regarded as reliable indicators of the perceived quality of the homework assigned by the teacher, we finally specified two multilevel models examining the effects of homework quality on student reports about homework effort. In both models, stable person characteristics were included as grand-mean centered covariates (gender, basic cognitive abilities, and the personality characteristic conscientiousness) at the individual level.[6] It is important to note that these covariates should be centered at the grand mean because interindividual differences in these covariates should be controlled for at the individual as well as the class level. Group-mean centering of these covariates would imply that we only control for individual differences in these covariates that persist within classes. Students' reports of the quality of their homework assignments were entered as individual variables (class-mean centered in model 1, grand-mean centered in model 2) and as class-average perceptions.

In the first model, individual student ratings were centered at their corresponding class mean (see Table 1). As mentioned above, the first model examines whether (a) individual perceptions of homework quality predict students' effort and (b) irrespective of individual perceptions, students' mean ratings of homework quality predict their effort (i.e., not partialling out individual student perceptions, but controlling for the other grand-mean centered covariates). The intercept $\hat{\gamma}_{00}$ is an adjusted overall mean for student self-reported homework effort after controlling for the covariates and is close to zero due to the standardization of the variables included. More interestingly, a statistically significant positive effect of student perceptions of homework quality on students' homework effort was found at both the individual ($\hat{\gamma}_{10} = .24$) and the class level ($\hat{\gamma}_{01} = .43$). Classes in which students collectively rated the homework assigned by their teacher to be of high quality showed higher average levels of homework effort than classes in which students rated the homework assigned to be of low quality. At the student level, there was also a positive effect of individual students' ratings of their homework quality on homework effort, indicating that students most likely to perceive their homework quality as high also reported high levels of homework effort. In addition, the (grand-mean centered) covariates indicated a statistically significant positive effect of conscientiousness and a negative effect of gender.

What would have happened if we had used grand-mean centering of the individual student ratings of homework quality? As pointed out above, a model that uses grand-mean centering (of the individual ratings) would examine whether (a) individual perceptions of homework quality predict effort and (b) controlling for individual students' perceptions, mean ratings of homework quality predict effort (i.e., partialling out individual students' perceptions and the grand-mean centered covariates). As expected, we obtained exactly the same regression coefficient at the individual level ($\hat{\gamma}_{10} = .24$), but a substantially lower effect at the class level ($\hat{\gamma}_{01} = .19$). As shown in detail in the Appendix, the between-class regression coefficient for homework quality in the CGM model is easily obtained from the CWC model. Because, in the CGM model, the effect of mean homework quality is adjusted for interindividual differences in student ratings, subtracting the value of the within-class regression coefficient of the CWC model, .24, from the value of the corresponding between-class coefficient, .43, yields the between-class partial regression coefficient of the CGM model, .19. The between-class regression coefficient is smaller when the level-1 predictor is centered at its grand mean because the effect of the class variable is reduced for interindividual differences in students' ratings. Hence, the partial between-class regression coefficient from the CGM model leads to a different interpretation of the effects of homework quality at the class level.[7] Now the central question is which between-class coefficient should be interpreted: the one from the CWC model or the one from CGM model? In the following, we would like to assist researchers in making the centering decision by giving reasons for choosing grand-mean or group-mean centering of individual student ratings.

5.1. Using grand-mean centering of individual student ratings

The standard textbook example for using grand-mean centering of individual level-predictor variables in multilevel modeling relates to the question of whether student socio-economic status (SES) affects student achievement (Hox, 2002; Raudenbush & Bryk, 2002; Snijders & Bosker, 1999). This standard example tests whether—in addition to an effect at the individual level—class- or school-aggregated SES is associated with student achievement. A statistically significant effect at the class or school level would support the notion that, even when individual students' SES is controlled, the composition of the learning group in terms of SES affects student outcomes and that the "context" matters. In this application of multilevel modeling, CGM needs to be applied to student SES at the individual level because we are interested in the partial effect of class- or school-aggregated SES (i.e., partialling out individual SES).[8] Why is CGM the correct centering option in this example? The contextual effect of SES on student achievement can be regarded as the net effect of a group variable when the effect of the same variable at the individual level is controlled (Raudenbush & Bryk, 2002). In the case of SES, it is reasonable to adjust for interindividual differences in SES because, to some extent, students attend different schools depending on their social background. If individual students' SES were not controlled—i.e., if CWC were used or individual SES were omitted from the model—an effect of the school-aggregated SES might be attributable to students' assignment to different schools depending on their individual SES.

Applying the same rationale to the analysis of student ratings of learning environments, use of CGM would imply that researchers wish to control for individual differences in students' perceptions

[5] The between-class variance was statistically significantly different from zero, $\tau^2 = .177$, SE = .035, $p < .001$.

[6] To enhance the interpretation of regression coefficients, all continuous variables were standardized (z-score with $M = 0$, $SD = 1$) at the individual level. Individual students' ratings of homework quality were aggregated but not re-standardized at the class level (thus, class-level effects are measured in terms of individual student level SDs; see Marsh & Roche, 1996).

[7] In the previous analyses, we did not control for group composition in terms of gender, conscientiousness, and basic cognitive abilities. Thus, an additional model was run in which classroom means of gender, conscientiousness, and basic cognitive abilities were included as additional predictor variables at level 2. None of these classroom means were statistically significantly related to the students' effort. The effect of the aggregated student ratings of homework quality remained almost unchanged ($\hat{\gamma}_{01} = .21$, SE = .08, $p < .01$).

[8] More precisely, if group-mean centering is used, the individual-level regression coefficient for SES needs to be subtracted from the between-group coefficient to give the partial effect between-group regression coefficient.

when assessing the effects of aggregated, collective perceptions. This might be a reasonable decision if there is reason to believe that students' assignment to different classes is associated with their tendency to perceive the classroom environment in a more or less positive light or with variables related to such a tendency. Such a rating tendency or rater bias (see Hoyt, 2000) typically involves a tendency to perceive a target as generally high or low on a rating scale. For example, it seems reasonable to suggest that high achieving students tend to perceive the quality of their homework in a more positive light and that these better students are more likely assigned to certain classrooms (e.g., through ability grouping). If these differences in prior ability are not controlled, using grand-mean centering of individual student ratings could adjust for individual differences in this rating tendency in student perceptions. Hence, the effect of the aggregated student ratings of homework quality might be overestimated if we did not control for this rater bias in student perceptions. On the other hand, there are also reasons that justify group-mean centering of student ratings. From our perspective, the most critical point, as explained in the next section, is that the rationale that underlies the aggregation of student perceptions of their learning environment is completely different from the rationale that underlies the aggregation of SES.

5.2. Using group-mean centering of individual student ratings

As pointed out above, when applying CWC, a researcher examines whether individual and collective perceptions of homework quality predict students' effort, but does not wish to control for differences in individual perceptions when assessing the effect of collectively perceived homework quality. What could be the rationale behind CGM of individual student ratings?

When student ratings are used to assess features of learning environments reflecting instructional or teacher variables, all individual perceptions are supposed to measure the same construct at the class level (e.g., quality of the homework assigned by the teacher in a class). From a measurement perspective, individual students act as interchangeable observers of a construct that exists at the class level and reflects a specific characteristic of the class, namely the quality of homework. It can be argued that this class-level construct affects students' perceptions (Cronbach, 1976; Miller & Murdock, 2007): an individual student's rating of the homework quality in his or her class depends on the actual quality of homework in that class. In this respect, student ratings of classroom features are completely different from a construct such as SES, which is used as a standard example for grand-mean centering. In the case of SES, the group-level variable is a summary index of a level-1 construct (e.g., individual students' SES) that is aggregated to the group level (e.g., school-average SES). The individual student scores are not interchangeable observations of a group construct, as are individual perceptions of homework quality. Rather, they are used to form the corresponding construct at the group level. Hence, it makes no sense to argue that the individual SES scores are affected by a corresponding group-level construct (i.e., school-average SES), because the individual SES values cause and do not reflect school-average SES. In other words, school-average SES does not exist independently of individual students' SES; however, a certain classroom feature can prevail in a classroom, even if individual students do not perceive it.

Applying this reasoning to the multilevel analysis of the effects of learning environments using student ratings, it could be argued that centering the individual perceptions at their group mean is preferable to grand-mean centering. As pointed out above, CGM adjusts for interindividual differences in students' perceptions of the learning environment and results in a partial regression coefficient at the group level. This adjustment might be questioned when a class-level construct such as homework quality that causes the individual perceptions is assessed, however. It can be shown (see Eq. (A.5) in the Appendix) that, in CGM, group differences in student outcomes are adjusted for group differences in student perceptions of the learning environment by using the relationship between the outcome measure and the student perceptions at the individual level. In the case of SES, it is reasonable to adjust for interindividual differences in SES because students are sorted into different schools depending on their social background. However, in the case of student ratings of classroom features, these between-class differences in individual students' perceptions of their learning environment can already be regarded as an effect of the group construct (e.g., homework quality) that is assessed by using individual students as informants on their learning environment. Thus, it could be argued that controlling for these differences would eliminate an essential component of the aggregated student perceptions. In contrast, no essential part of aggregated SES is eliminated when interindividual differences in SES are controlled because (unlike the case of homework quality) an individual's SES is not the result of observing a school-average SES.

To sum up, centering student ratings of their learning environment is a complex issue in MLM. We have shown that conclusions about the effects of learning environments can be influenced by the choice of a specific centering option for the student ratings. In addition, we have given reasons for selecting either grand-mean centering or group-mean centering of student ratings. However, we believe that in research practice it is often difficult to make a clear-cut decision for one of these options. Hence, it is important that researchers always specify which centering option was chosen for their predictor variables (see also Ferron et al., 2008) so that other researchers can infer (using the relation between CWC and CGM given in the Appendix) how the results might differ if the other centering option had been chosen.

6. Summary and conclusion

This article addressed conceptual and methodological issues that arise when student ratings are used to assess features of learning environments. Three crucial elements were discussed: (1) Researchers who are interested in identifying effects of learning environments should focus on ratings that are aggregated at the relevant level (e.g., class or school level). (2) It is imperative to assess the reliability of aggregated student ratings before these perceptions are related to the outcome variables. (3) When conducting multilevel analyses, researchers need to choose an appropriate centering option; when individual and aggregated student ratings are both included as predictor variables, effects of the learning environment can be substantially affected by choosing either grand-mean or group-mean centering.

In research on learning environments, students are often used as informants on different aspects of the learning environment. Frequently, only Cronbach's α is reported to document the reliability of student perceptions. Cronbach's α assesses the internal consistency of a person measure; however, the reliability of a classroom-level construct such as homework quality depends on the intraclass correlation and the number of students rating the group-level construct under investigation. In our application, we demonstrated that even with an $ICC(1)$ of modest magnitude, satisfactory reliability of the average student ratings could be reached with an average cluster size of 15–20 students per class. However, recent developments in multilevel modeling make it possible to correct for the unreliable assessment of the group mean when estimating group effects (Lüdtke et al., 2008; Rabe-Hesketh, Skrondal, & Pickles, 2004).

In the present study, we focused on students' ratings of features of their learning environment. Of course, other data sources capture different aspects of teaching and learning processes in the classroom. Previous research has demonstrated the differential validity of these different data sources when it comes to explaining student learning (Clausen, 2002; Urdan, 2004). As emphasized by Staub (2007), for example, video recordings capture teaching in real-life classroom settings and allow the systematic analysis of specific instances of certain practices ("low inference" measures; see Walberg & Hartel, 1980). To make recommendations for classroom practice, researchers must know what actually goes on during lessons. Consequently, future studies assessing the impact of learning environments on students' outcomes would be well advised to use a mix of data sources, rather than a single-method approach (Raudenbush, 2005; Seidel & Shavelson, 2007).

Several important issues relating to the use of student ratings were not addressed in our article. First, there is consensus among educational researchers that multilevel modeling is the most appropriate method for analyzing the effects of student perceptions of learning environments on student outcomes. However, researchers still have to make the crucial decision of how to center the student ratings at the individual level. In the present article, we argued that CWC can be justified as an appropriate centering option when students' ratings of their learning environment are included in multilevel models as independent variables. Our central argument for choosing CWC is that constructs that assess different aspects of students' learning environments are generic group-level constructs that affect individual perceptions, and that using CGM to control for interindividual differences in these ratings would eliminate an essential component of the aggregated ratings. At the same time, we also presented reasons for favoring CGM as a centering option. The central argument was that differences in student ratings of the learning environment could, to a certain degree, be the effect of rater bias and that CGM might be justified because it controls for that rater bias. When the independent variable, students' ratings of their learning environment, and the outcome variable, are all assessed using the same method (e.g., self report), it is possible that correlated method bias might distort the within-class regression coefficient as well as the between-class regression coefficient. Given that little is known about the magnitude of correlated method bias of rating data (Hoyt, 2000), more research is clearly needed into the role of rater- and method-specific effects in multilevel analysis of students' ratings of learning environments and into how far CGM is able to control for possible effects of rater bias.

Second, in our discussion of centering in multilevel models, we focused on models that allow the intercept to vary between classes only (random-intercept model). Frequently, researchers are interested in group effects that modify not only the mean level of the outcome, but also the relationship between variables within groups (cross-level interaction). For instance, the relationship between a student's rating of the social climate and his/her positive emotions might be assumed to be affected by certain other features of the classroom environment (e.g., the gender of a teacher, competitiveness in class). In this case, however, there are clear recommendations in the multilevel modeling literature stating that CWC of the predictor at the individual level should be used to identify the cross-level interaction effect (for further details, see Enders & Tofighi, 2007; Hofmann & Gavin, 1998; Raudenbush & Bryk, 2002).

Third, the *ICC*(2) estimates the reliability with which the aggregated student ratings differentiate between classes or schools. Of course, calculating the *ICC*(1) and *ICC*(2) is only a first step—albeit a very important one—toward adopting a multilevel perspective in assessing the psychometric properties of students' ratings. The logical extension of a (univariate) analysis of homogeneity and reliability would be a (multivariate) examination of the factor structure when students rate several aspects of their learning environment, such as multiple classroom goal structures (e.g., mastery goals, performance goal orientation; Karabenick, 2004) or different instructional features. (e.g., homework quality, homework control; Trautwein et al., 2006) When multilevel exploratory factor analysis (see Reise, Ventura, Nuechterlein, & Kim, 2005) and multilevel confirmatory factor analysis (Mehta & Neale, 2005; Rabe-Hesketh et al., 2004) are used, it is important to determine the extent of within- and between-class differences in the factor structure of student ratings.

In conclusion, research using student ratings to analyze the effects of learning environments entails some major methodological challenges. Researchers examining differences between learning environments often have no choice but to use aggregated student ratings of the classroom or school environment. This requires considerable sample sizes at the group level, and makes studies that seek to analyze the effects of features of learning environments rather cost-intensive. We can only agree with Cronbach, who drew attention to this problem of social scientific research: "A social science must deal with collectives, and the cost of obtaining data on collectives is great" (Cronbach, 1976, p. 10.12).

Appendix A

A.1. The basic multilevel model to analyze effects of students' ratings

In this section, we briefly describe the basic multilevel model used to analyze effects of students' ratings of their learning environment on student outcomes. In so doing, we adopt the terminology used by Enders and Tofighi (2007); (see also Hofmann & Gavin, 1998; Kreft et al., 1995; Raudenbush & Bryk, 2002). We will show that the two forms of centering—centering at the group mean (CWC) and centering at the grand mean (CGM)—produce estimates for the effects of aggregated student ratings that differ in value, but also in meaning. The research question chosen from classroom climate research is whether students' positive emotions (*EMOTION*) in class are affected by their perceptions of the social climate (*CLIMATE*; e.g., student/teacher relations). Let us assume a two-level structure, with students nested within classes. Both individual students' perceptions of the social climate and class-average perceptions are included as predictor variables.

Applying the MLM notation used by Raudenbush and Bryk (2002) gives the following relation at the first level:

$$EMOTION_{ij} = \beta_{0j} + \beta_{1j}(CLIMATE_{ij} - \bar{x}_{CLIMATE_j}) + r_{ij} \quad (A.1)$$

where the variable $EMOTION_{ij}$ is the outcome for student i in class j predicted by the intercept β_{0j} of class j and the regression slope β_{1j} in class j. The predictor variable $CLIMATE_{ij}$ is centered at the respective class mean $\bar{x}_{CLIMATE_j}$ Thus, applying CWC to the individual student's climate perception gives his/her relative impression of the classroom climate within his/her class. Technically, CWC of the individual-level predictor yields an intercept β_{0j} equal to the expected value of $EMOTION_{ij}$ for a student whose value on $CLIMATE_{ij}$ is equal to the class mean $\bar{x}_{CLIMATE_j}$. From ordinary regression analysis we know that the prediction line for each group goes through the point corresponding to the mean value of the independent variable X and the mean value of the dependent variable Y for that group. It follows that the level-1 intercept β_{0j} in Eq. (A.1) is the unadjusted mean for class j (see Enders & Tofighi, 2007). At level-2, the level-1 intercepts β_{0j} and slopes β_{1j} are dependent variables:

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \gamma_{01}(\bar{x}_{CLIMATE_j}) + u_{0j} \\ \beta_{1j} &= \gamma_{10} \end{aligned} \quad (A.2)$$

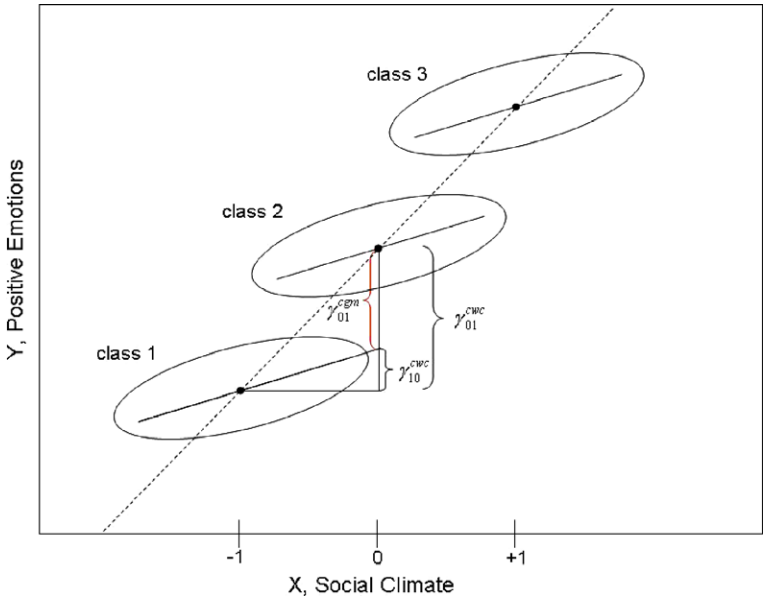

Fig. 4. Illustration of the difference between grand-mean centering and group-mean centering (see Raudenbush & Bryk, 2002, p. 140).

where γ_{00} and γ_{10} are the level-2 intercepts and γ_{01} is the slope relating $\bar{x}_{CLIMATE_j}$ to the intercepts β_{0j} (i.e., unadjusted class means $\bar{y}_{EMOTION_j}$) from the level-1 equation. As our example shows, only the level-1 intercepts have a level-2 residual u_{0j}. Multilevel models that allow the intercept to deviate from its predicted value are also called random-intercept models (e.g., Raudenbush & Bryk, 2002). Note that in these models, group effects are only allowed to modify the mean level of the outcome for the group. The distribution of effects among persons within groups (e.g., slopes β_{1j}) is left unchanged. Inserting the level-2 equations into the level-1 equation gives:

$$EMOTION_{ij} = \gamma_{00} + \gamma_{10}(CLIMATE_{ij} - \bar{x}_{CLIMATE_j}) + \gamma_{01}(\bar{x}_{CLIMATE_j}) + u_{0j} + r_{ij} \quad (A.3)$$

Furthermore, it can now easily be seen that γ_{10} is the within-class regression coefficient describing the relationship between $EMOTION_{ij}$ and $(CLIMATE_{ij} - \bar{x}_{CLIMATE_j})$ within classes and that γ_{01} is the between-class regression coefficient denoting the relationship between unadjusted class means $\bar{y}_{EMOTION_j}$ and $\bar{x}_{CLIMATE_j}$ (Cronbach, 1976). Note that the two predictors $(CLIMATE_{ij} - \bar{x}_{CLIMATE_j})$ and $\bar{x}_{CLIMATE_j}$ are orthogonal (i.e., uncorrelated), because CWC purges the individual scores of all between-class mean differences. Hence, γ_{10} and γ_{01} quantify the independent association between positive emotions and students' ratings of the social climate at level-1 and level-2, respectively.

Instead of using group mean centering of the predictor variable—where the class mean of the level-1 predictor is subtracted from each case—researchers can also decide to center the individual student's rating at its grand mean. Substituting the class-mean $\bar{x}_{CLIMATE_j}$ in Eq. (A.1) by the grand-mean $\bar{x}_{CLIMATE}$ gives the following level-1 equation:

$$EMOTION_{ij} = \beta_{0j} + \beta_{1j}(CLIMATE_{ij} - \bar{x}_{CLIMATE}) + r_{ij} \quad (A.4)$$

To grasp the differences between CGM and CWC, it is important to note that the interpretation of the intercept β_{0j} has changed. In the case of CGM, the intercept β_{0j} is the predicted score of a student whose social climate rating is at the grand mean (i.e., when $CLIMATE_{ij} = \bar{x}_{CLIMATE}$, the predicted *EMOTION* score is β_{0j}). As can be shown, this CGM gives a completely different interpretation of the level-1 intercept β_{0j}. In the case of CWC, the level-1 intercept is the unadjusted mean for class *j*. Taking the expectations of Eq. (A.4) and rearranging the terms for CGM (see Enders & Tofighi, 2007; Raudenbush & Bryk, 2002) yields

$$\beta_{0j} = \mu_{EMOTION_j} - \beta_{1j}(\bar{x}_{CLIMATE_j} - \bar{x}_{CLIMATE}) \quad (A.5)$$

As can be seen from Eq. (A.5) for CGM, the level-1 intercept β_{0j} is equal to the average emotion for class *j*, minus an adjustment that depends on two factors: (1) the regression slope and (2) the deviation between the average climate for class *j* and the grand mean. Eq. (A.5) corresponds to the formula usually applied in ANCOVA to calculate means adjusted for interindividual differences in the covariate (Maxwell & Delaney, 2004). ANCOVA investigates what the class effect on positive emotions would have been if the classes had been at the same mean value on the individual students' perceptions of social climate. Thus, the level-1 intercepts β_{0j} in CGM are controlled for interindividual differences in the corresponding level-1 predictor. Note that the level-1 slope β_{1j} is used to adjust for differences between groups in mean social climate.

Further insights into the differences between CGM and CWC can be gained by looking at the combined model for the CGM case. Inserting Eq. (A.2) into Eq. (A.4) yields

$$EMOTION_{ij} = \gamma_{00} + \gamma_{10}(CLIMATE_{ij} - \bar{x}_{CLIMATE}) + \gamma_{01}(\bar{x}_{CLIMATE_j}) + u_{0j} + r_{ij} \quad (A.6)$$

In contrast to the group-mean centered model, where the predictor variables are orthogonal, in this grand-mean centered model the predictors $(CLIMATE_{ij} - \bar{x}_{CLIMATE})$ and $\bar{x}_{CLIMATE_j}$ are no longer independent. This is because in CGM students' deviations from the grand mean $(CLIMATE_{ij} - \bar{x}_{CLIMATE})$ also include the student's class deviation from the grand mean $(\bar{x}_{CLIMATE_j} - \bar{x}_{CLIMATE})$. Hence, as shown by Eq. (A.5), $\bar{x}_{CLIMATE}$ predicts the adjusted means and γ_{01} is a partial regression coefficient.

At a casual glance, the CGM and CWC models would seem to be completely different. However, a simple relationship holds between the regression coefficients of the CGM and CWC model. For the fixed effects, the following relation holds for the level-2 between-class regression coefficients:

$$\gamma_{01}^{cgm} = \gamma_{01}^{cwc} - \gamma_{10}^{cwc} \quad (A.7)$$

The within-group regression coefficient at level-1 will be the same in both models: $\gamma_{10}^{cgm} = \gamma_{10}^{cwc}$. Hence, the results for the fixed part of the grand-mean centered model can be obtained from the group-mean centered model by a simple subtraction. In fact, these models are mathematically equivalent (i.e., they generate the same set of means and dispersion; see Kreft et al., 1995). It is important to add that mathematical equivalence only holds for the random-intercept model. If the slopes (i.e., the regression coefficients within groups) vary across groups, the variance components are no longer equivalent (see Kreft et al., 1995).

Fig. 4 illustrates the difference between γ_{01}^{cgm} and γ_{01}^{cwc} to show how the centering decision affects the interpretation of the parameters. The diagram is adapted from Raudenbush and Bryk (2002, p. 140), who originally used it to clarify the interpretation of a contextual effect. It displays three school classes that differ from one another by one unit in their mean social climate. The within-class relationship between individual students' ratings of social climate and positive emotions, γ_{10}^{cwc}, is plotted for each class. For class 1, γ_{10}^{cwc} is the expected difference in positive emotions between two students in the same class whose ratings of the social climate differ by one unit. Similarly, γ_{01}^{cwc} represents the expected difference between the means of two classes (class 1 versus class 2) that differ by one unit in mean social climate. Thus, the adjusted group effect of social climate from the grand-mean centered model, $\gamma_{01}^{cgm} = \gamma_{01}^{cwc} - \gamma_{10}^{cwc}$, is now the expected difference in positive emotions between two students who have the same individual perception of the social climate, but who attend different classes that differ by one unit in mean social climate.

References

Aleamoni, L. M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education, 13*, 153–166.

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

Anderson, G. J. (1973). *The assessment of learning environments: A manual for the LEI and MCI*. Halifax: Atlantic Institute for Education.

Anderson, C. S. (1982). The search for school climate: A review of the research. *Review of Educational Research, 52*, 368–420.

Anderson, L. W., Ryan, D., & Shapiro, B. (Eds.). (1989). *The IEA classroom environment study*. New York: Pergamon Press.

Bickel, R. (2007). *Multilevel analysis for applied research*. New York: Guilford Press.

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.

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

Church, M. A., Elliot, A. J., & Gable, S. L. (2001). Perceptions of classroom environment, achievement goals, and achievement outcomes. *Journal of Educational Psychology, 93*, 43–54.

Clausen, M. (2002). *Unterrichtsqualität: Eine Frage der Perspektive? [Instructional quality: A matter of perspective?]*. Münster: Waxmann.

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

de Corte, E. (2001). Technology-supported learning environments. In N. J. Smelser & P. B. Baltes (Eds.). *International encyclopedia of the social and behavioral sciences* (Vol. 23, pp. 15527–15532). Oxford, UK: Elsevier.

De Jong, R., & Westerhof, K. J. (2001). The quality of student ratings of teacher behaviour. *Learning Environments Research, 4*, 51–85.

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.

Ferron, J., Hogarty, K. Y., Dedrick, R. F., Hess, M. R., Niles, J. D., & Kronrey, J. D. (2008). Reporting results from multilevel analyses. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 391–426). Charlotte, NC: Information Age Publishing.

Fraser, B. J. (1980). *Criterion validity of an individualized classroom environment questionnaire*. Sydney: McQuaire University.

Fraser, B. J. (1991). Two decades of classroom environment research. In H. J. Walberg (Ed.), *Educational environments: Evaluation, antecedents and consequences* (pp. 3–27). Elmsford, NY: Pergamon Press.

Fraser, B. J., & Fisher, D. L. (1982). Predicting students' outcomes from their perceptions of classroom psycho-social environment. *American Educational Research Journal, 19*, 498–518.

Frenzel, A. C., Pekrun, R., & Goetz, T. (2007). Perceived learning environments and students' emotional experiences: A multilevel analysis of mathematics classrooms. *Learning and Instruction, 17*, 478–493.

Friedel, J. M., Cortina, K. S., Turner, J. C., & Midgley, C. (2007). Achievement goals, efficacy beliefs and coping strategies in mathematics: The roles of perceived parent and teacher goal emphasis. *Contemporary Educational Psychology, 32*, 434–458.

Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist, 52*, 1182–1186.

Hofmann, D. A., & Gavin, M. B. (1998). Centering decisions in hierarchical linear models: Implications for research in organizations. *Journal of Management, 24*, 623–641.

Hox, J. J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods, 5*, 64–86.

James, L. R. (1982). Aggregation bias in estimates of perceptual agreement. *Journal of Applied Psychology, 67*, 219–229.

Kamata, A., Bauer, D. J., & Miyazaki, Y. (2008). Multilevel measurement modeling. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 345–388). Charlotte, NC: Information Age Publishing.

Kane, M. T., & Brennan, R. L. (1977). The generalizability of class means. *Review of Educational Research, 47*, 267–292.

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, A., Gheen, M., & Midgley, C. (2002). Classroom goal structure and student disruptive behaviour. *British Journal of Educational Psychology, 72*, 191–211.

Karabenick, S. A. (2004). Perceived achievement goal structure and college student help seeking. *Journal of Educational Psychology, 96*, 569–581.

Kreft, I., & de Leeuw, J. (1998). *Introducing multilevel modeling*. London: Sage.

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.

Kunter, M., Baumert, J., & Köller, O. (2007). Effective classroom management and the development of subject-related interest. *Learning and Instruction, 17*, 494–509.

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

LeBreton, J. M., & Senter, J. L. (2008). Answers to twenty questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11*, 815–852.

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., 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., Trautwein, U., Kunter, M., & Baumert, J. (2006). Reliability and agreement of student ratings of the classroom environment – A reanalysis of TIMSS data. *Learning Environments Research, 9*, 215–230.

Marsh, H. W., & Bailey, M. (1993). Multidimensional students' evaluations of teaching effectiveness: A profile analysis. *The Journal of Higher Education, 64*, 1–18.

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist, 52*, 1187–1197.

Marsh, H. W., & Roche, K. J. (1996). The negative effects of school-average ability on academic self-concept: An application of multilevel modeling. *Australian Journal of Education, 40*, 65–87.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Erlbaum.

Mayer, D. P. (1999). Measuring instructional practice: Can policy makers trust survey data? *Educational Evaluation and Policy Analysis, 21*, 29–45.

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

Midgley, C., Maehr, M. L., Hruda, L. A., Anderman, E., Anderman, L., Gheen, M., et al. (2000). *Manual for the patterns of adaptive learning scale (PALS)*. Ann Arbor: University of Michigan.

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

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

Moos, R. H. (1979). *Evaluating educational environments*. San Francisco: Jossey-Bass.

Moos, R. H., & Trickett, E. J. (1974). *Classroom environment scale manual*. Palo Alto: Consulting Psychologists Press.

Murray, H. A. (1938). *Explorations in personality*. New York: Oxford University Press.

OECD. (2001). Knowledge and skills for life. First results from the OECD Programme for international student assessment (PISA) 2000. OECD: Author.

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.

Peugh, J. L., & Enders, C. K. (2005). Using the SPSS mixed procedure to fit cross-sectional and longitudinal multilevel models. *Educational and Psychological Measurement, 65*, 717–741.

Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher, 31*, 3–14.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika, 69*, 167–190.

Raudenbush, S. W. (2005). Learning from attempts to improve schooling: The contribution of methodological diversity. *Educational Researcher, 34*, 25–31.

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

Reise, S. P., Ventura, J., Nuechterlein, K. H., & Kim, K. H. (2005). An illustration of multilevel factor analysis. *Journal of Personality Assessment, 84*, 126–136.

Ryan, A. M., Gheen, M. H., & Midgley, C. (1998). Why do some students avoid asking for help? An examination of the interplay among students' academic efficacy, teachers' social-emotional role, and the classroom goal structure. *Journal of Educational Psychology, 90*, 528–535.

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

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

Spybook, J. (2008). Power, sample size, and design. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 273–314). Charlotte, NC: Information Age Publishing.

Staub, F. C. (2007). Mathematics classroom cultures: Methodological and theoretical issues. *International Journal of Educational Research, 46*, 319–326.

Stern, G. G. (1970). *People in context. Measuring person–environment congruence in education and industry*. New York: John Wiley.

Trautwein, U., Lüdtke, O., Schnyder, I., & Niggli, A. (2006). Predicting homework effort: Support for a domain-specific, multilevel homework model. *Journal of Educational Psychology, 98*, 438–456.

Trouilloud, D., Sarrazin, P., Bressoux, P., & Bois, J. (2006). Relation between teachers' early expectations and students' later perceived competence in physical education classes: Autonomy-supportive climate as a moderator. *Journal of Educational Psychology, 98*, 75–86.

Turner, J. C., & Meyer, D. K. (2000). Studying and understanding the instructional contexts of classrooms: Using our past to forge our future. *Educational Psychologist, 35*, 69–85.

Turner, J. C., Midgley, C., Meyer, D. K., Gheen, M., Anderman, E. M., Kang, Y., et al. (2002). The classroom environment and students' reports of avoidance strategies in mathematics: A multimethod study. *Journal of Educational Psychology, 94*, 88–106.

Urdan, T. (2004). Using multiple methods to assess students' perceptions of classroom goal structures. *European Psychologist, 9*, 222–231.

Urdan, T., Midgley, C., & Anderman, E. M. (1998). The role of classroom goal structure in students' use of self-handicapping strategies. *American Educational Research Journal, 35*, 101–122.

Walberg, H. J. (1972). Social environment and individual learning: A test of the Bloom model. *Journal of Educational Psychology, 63*, 69–73.

Walberg, H. J., & Hartel, G. D. (1980). Validity and use of educational environment assessments. *Studies in Educational Evaluation, 6*, 225–238.

Wendorf, C. A., & Alexander, S. (2005). The influence of individual- and class-level fairness-related perceptions on student satisfaction. *Contemporary Educational Psychology, 30*, 190–206.

Wong, A. F., Young, D. J., & Fraser, B. J. (1997). A multilevel analysis of learning environments and student attitudes. *Educational Psychology, 17*, 449–468.

Wubbels, T., Brekelmans, M., & Hooymayers, H. P. (1992). Do teacher ideals distort the self-reports of their interpersonal behavior? *Teaching and Teacher Education, 8*, 47–58.