Journal of Educational Psychology
2014, Vol. 106, No. 1, 289–300

0022-0663/14/$12.00 DOI: 10.1037/a0033829

Classroom Composition, Classroom Management, and the Relationship Between Student Attributes and Grades

Jan Hochweber
German Institute for International Educational Research, Frankfurt, Germany

Ingmar Hosenfeld
University of Koblenz-Landau

Eckhard Klieme
German Institute for International Educational Research, Frankfurt, Germany

The present study examined the extent to which the relationships between student self-reported math grades and different types of student variables (standardized math test scores, interest and effort in math, parental education) are predicted by classroom composition and teachers' classroom management. Based on a representative sample of 31,038 8th-grade students from 1,470 classrooms, multilevel regression analyses revealed that grades were less strongly related to students' test scores and more strongly related to students' effort in classrooms with an unfavorable academic composition (i.e., low average test performance). Classroom management was found to moderate the association between academic classroom composition and the parental education–grade relationship, indicating a noticeable grade advantage of students with high parental education in classrooms with both an unfavorable academic composition and ineffective classroom management. Our findings highlight the relevance of classroom composition and classroom management to research on teachers' grading and point toward possible ways to improve current grading practices.

Keywords: achievement, classroom composition, classroom management, grading

Teacher-assigned grades represent a composite of various student-related attributes. Teachers take into account a considerable number of factors in determining grades. While academic achievement appears to be typically most relevant (Brookhart, 1994; McMillan, 2001), many teachers also give substantial weight to nonachievement factors such as effort and ability (Cross & Frary, 1999; McMillan, 2001). Not surprisingly, then, teacher-assigned grades are related to academic knowledge as assessed by standardized tests and, over and above test scores, to various noncognitive outcomes. While test scores traditionally account for 25% to 35% of variance in grades (Bowers, 2011), the remaining variance seems to be largely explained by student behaviors such as participating in class and doing the work assigned (Willingham, Pollack, & Lewis, 2002). There is, however, yet another source of variance that deserves attention: Student background variables such as gender and socioeconomic status (SES) can be related to grades over and above test scores, prior grades, and noncognitive outcomes (Bennett, Gottesman, Rock, & Cerullo, 1993; Jussim, 1989).

To date, research has focused on the overall relationship between student attributes and grades. Few studies have distinguished between student and class or school level variance in grades and have related them to class or school level variables. These studies have been typically concerned with between-class or between-school differences in grading levels (Klapp Lekholm, 2011; Klapp Lekholm & Cliffordson, 2008), while variation in the relationship between student attributes and grades has hardly been investigated. Surveys on teachers' grading practices have focused on the impact of teacher characteristics (Bonner & Chen, 2009; McMillan, 2001), while research on the role of classroom and school context has been scarce. Very few studies have used actual grades to investigate the impact of class or school variables on the student attribute–grade relationship but have been based on small and probably nonrepresentative samples. In addition, the available studies have not considered the relationship between grades and student background variables but have exclusively focused on test scores and noncognitive outcomes.

This scarcity of research seems surprising, as the extent to which student attributes are related to grades is important from several perspectives. From the perspective of instruction, teachers' use of nonachievement factors in grading has been interpreted as an attempt to serve important goals of instruction, including motivating students and providing disadvantaged students with a fair chance to succeed (Brookhart, 1993; McMillan, 2001). Grades can also be used to

This article was published Online First August 12, 2013.

Jan Hochweber, Department of Educational Quality and Evaluation, German Institute for International Educational Research, Frankfurt, Germany; Ingmar Hosenfeld, Centre for Educational Research, University of Koblenz-Landau, Landau, Germany; Eckhard Klieme, Department of Educational Quality and Evaluation, German Institute for International Educational Research.

We thank Benjamin Fauth and Alexander Naumann for their helpful comments.

Correspondence concerning this article should be addressed to Jan Hochweber, Deutsches Institut für Internationale Pädagogische Forschung, Schloßstraße 29, 60486 Frankfurt am Main, Germany. E-mail: hochweber@dipf.de

ensure student cooperation, that is, as an element of classroom management (Brookhart, 1993, 1994). Thus, investigating systematic differences in the relationship of noncognitive outcomes and grades may provide insights in how teachers use grades as a tool of instruction (Rakoczy, Klieme, Bürgermeister, & Harks, 2008). From the educational career perspective, the factors influencing grades are important because of the implications for students' future prospects. Grades serve as selection and certification criteria but also guide students and parents in identifying areas of special ability (Hallinan, 1992; Lysne, 1984). If grades are biased against members of a specific social group, this may impair these students' educational career chances. Thus, investigating systematic differences in the relationship between student background variables and grades may help to ensure fairness in grade-based decisions.

The present study focuses on the relation of two types of variables, namely, (a) the student composition of the class and (b) teachers' classroom management, to the relationship between student attributes and grades. Composition variables have been found related to the grading level of classrooms and schools (Klapp Lekholm & Cliffordson, 2008), but their impact on the student attribute–grade relationship has, to our knowledge, not yet been considered. To help fill this research gap, we examined the association of student self-reported report card math grades with standardized math test scores, two noncognitive outcomes (interest and effort in math), and one background variable (parental education) in classes differing with respect to their academic, motivational, and social student composition.

Many practicing teachers seem to regard grading as an important aspect of classroom management (Brookhart, 1993, 1994). Nevertheless, classroom management has been hardly considered in the literature on assessment and grading (Brookhart, 2004). One way in which teachers' classroom management skills might become relevant to grading is by influencing the way in which teachers react to specific classroom conditions. Based on this notion, we investigated to which extent teachers' classroom management moderates the association between the classroom composition variables and the student attribute–grade relationships.

Since grades are assigned by teachers and depend on their cognitions and decisions related to grading, any influence of classroom composition on the relationship between student attributes and grades must be mediated by teacher-related variables. Drawing on studies from several lines of research, we argue that classroom composition is relevant to the relationships between student attributes and grades as it affects teachers' grading practices, judgment accuracy, and susceptibility to bias. Furthermore, we suggest that classroom management affects the attribute–grade relationships by moderating the impact of classroom composition on teachers' grading practices, judgment accuracy, and susceptibility to bias.

Teachers' Grading Practices

Some evidence on the role of classroom composition comes from a teacher survey by McMillan (2001). Teachers emphasized academic achievement as a grading factor most in advanced/advanced placement (AP) classes, less in standard classes, and least in basic classes, while the grading of homework and the use of "nontest" indicators in borderline cases were more prevalent in standard and basic classes than in advanced/AP classes. According to McMillan (2001), such practices might "make it easier for teachers to give passing grades" (p. 31). Another way to interpret them is as a means to establish control in "difficult" classes. Rakoczy et al. (2008) reported that student involvement contributed stronger to math grades in classes with a negative student–teacher relationship. Ennis (1995) found that teachers in urban high schools with a high percentage of ethnic minority students used grades to "bribe" students to spend a minimum of effort. In a study by Howley, Kusimo, and Parrott (2000), teachers in troubled schools tended toward an "ethos of effort," considering effort strongly in their grading. Taken together, these studies suggest that teachers in high-ability classes put emphasis on academic achievement rather than nonachievement factors, while teachers in classes with low levels of student ability and motivation put more weight on nonachievement factors and less on academic achievement.

Accuracy of Teacher Judgments of Student Achievement

Most research on teachers' judgment accuracy has been concerned with teachers' ability to predict the rank order of student test scores in their class. Only one study has, to our knowledge, considered classroom composition to explain variation in this ability. Martínez, Stecher, and Borko (2009) reported lower correlations between teacher ratings and student test scores in classes with a higher percentage of nonnative English speakers. According to Martínez et al., this may have resulted from the teachers' different familiarity with this group of students and problems involved in the assessment of these students or, alternatively, from "added demands that these classrooms place on teachers [which] could result in less consistent or meticulous student evaluations" (Martínez et al., 2009, p. 96). They also speculated that judgment accuracy might depend on classroom variables such as the prevalence of conduct problems.

Some support for this notion comes from Funder's Realistic Accuracy Model (RAM; Funder, 1995). In RAM, judgment accuracy depends on the relevance, availability, detection, and utilization of behavioral cues. Letzring, Wells, and Funder (2006) found that subjects who interacted for longer time periods and in situations where more personality-related information was available showed higher levels of realistic accuracy. Consequently, for a teacher's judgments to be accurate, achievement-related cues must be produced in the first place (relevance) and be perceivable to the teacher (availability). In classes with an unfavorable student composition, disciplinary problems are more prevalent (Barth, Dunlap, Dane, Lochman, & Wells, 2004; Rindermann, 2007), and teachers should be involved more in management activities and less in achievement-related interactions, which might impair their ability to profoundly assess student achievement. If so, the relationship between cognitive outcomes and grades should be lower in classes with an unfavorable student composition.[1]

Perceptual Bias in Teacher Judgments

Student background variables' effects on grades have been attributed to perceptual bias, resulting from inaccurate teacher expectancies. Naturally, this bias disadvantages groups for which

[1] Classes with low levels of student achievement or motivation or with a large proportion of (e.g.) low-SES students are referred to as classes with an "unfavorable" composition. This expression is not meant pejorative but only serves as a concise term for these types of classes.

teachers tend to hold low expectancies, like students with low SES and from certain ethnic minorities (Dusek & Joseph, 1983). Expectancy-based bias can be of considerable size, among others, in specific judgmental situations. Particularly relevant in the present context is the situation's cognitive demands. Correct inference of a target person's characteristics depends crucially on the perceiver's cognitive resources (Fiske & Neuberg, 1990; Kruglanski, Pierro, Mannetti, Erb, & Chun, 2007). Subjects under high cognitive load tend to produce biased inferences (Macrae, Hewstone, & Griffiths, 1993) and to be less likely to correct erroneous judgments (Gilbert & Osborne, 1989). Kruglanski and Freund (1983), for example, found increased ethnic stereotyping in a grading task when cognitive resources were reduced due to time pressure (see also Blair & Banaji, 1996; Pratto & Bargh, 1991). Consequently, if a teacher is often busy with task-irrelevant activities, as would be expected in classes with an unfavorable composition, the ability to assess student achievement should be compromised, and grading should be more biased against low-expectancy social groups.

Compensation for Social Disadvantage

Contrary to the notion of expectancy-based bias, Martínez et al. (2009) found that teachers' achievement ratings were *higher* for poor and minority students than would have been expected from their test scores. A possible explanation is "that teachers compensate for perceived disadvantages faced by these groups by adjusting their ratings up—or, alternatively, adjusting their criteria and expectations down" (Martínez et al., 2009, p. 97). Teachers are known to care a lot about the social consequences of their grading, and many of them use different criteria for high- and low-performing students and feel as if they are their students' "advocates" (Brookhart, 1993). Such attitudes might support grading strategies that favor students from disadvantaged groups. However, teachers might feel a different need to compensate for social disadvantages in different classroom contexts. If a class is highly supportive toward learning, teachers might have less reason to compensate for social disparities by adjusting students' grades. If a class provides an adverse environment for learners, as often occurs in classes with an unfavorable student composition, compensation by adjusting grades might be more likely.

The Role of Teachers' Classroom Management

Teachers' ability to create a functioning learning environment depends crucially on her or his classroom management skills. Effective classroom management rests on two key principles, identifying desirable student behaviors and preventing undesirable ones (Emmer & Stough, 2001; Kunter, Baumert, & Köller, 2007). To identify desirable behaviors, teachers have to communicate clear rules and establish stable routines. To prevent disruptions and ensure effective time use, teachers have to monitor what is taking place in the classroom ("withitness"; Kounin, 1970) and intervene immediately and effectively if necessary.

Classroom management becomes particularly relevant in "difficult" classes. These classes require an adaptation of strategies to their social, behavioral, and academic context (Emmer & Stough, 2001; Evertson, 1982). Unfortunately, many teachers feel unprepared for dealing with these classes (Merrett & Wheldall, 1992) and are stressed and dissatisfied by student misbehavior (Punch & Tuettemann, 1990). Confronted with students unwilling to cooperate, teachers with management problems might be tempted to exchange good grades for cooperation or to use grades to enforce discipline (Cothran & Ennis, 1997; Ennis, 1995; Pace & Hemmings, 2007). Successful classroom managers might also use "grades as pay" (Brookhart, 1993) to some extent but should not have to rely excessively on grades to handle a demanding class. This suggests that teachers put particular emphasis on specific nonachievement factors (e.g., effort) in grading if they teach a class with an unfavorable student composition but lack classroom management skills.

Good classroom management optimizes learning time (LePage et al., 2005), and thus increases teachers' opportunities to monitor students' learning progress. Ineffectively managed classes, on the other hand, should compromise teachers' capability to accurately and unbiasedly judge student achievement by providing less opportunity to monitor learning progress and putting more cognitive demands on the teacher. This might be especially pronounced in "difficult" classes, which more likely produce interruptions and disciplinary problems (Barth et al., 2004; Roland & Galloway, 2002). Accordingly, an unfavorable classroom composition should particularly undermine teachers' ability to accurately and unbiasedly judge student achievement if combined with ineffective classroom management. That is, the relation between cognitive outcomes and grades should tend to be lower, and the relation between background variables and grades should tend to be stronger if teachers have a class with an unfavorable composition but lack classroom management skills.

The Present Study

We studied two research questions concerning the role of classroom composition and classroom management in moderating the relationship between student attributes and grades. The first research question addressed the power of classroom composition variables to predict the relationship between four student attributes (math test scores, interest and effort in math, parental education) and student self-reported math grades. In line with previous research (e.g., Opdenakker, Van Damme, De Fraine, Van Landeghem, & Onghena, 2002), classroom composition was defined in terms of student attributes aggregated to the class level. Class-average math test performance, interest in math, and parental education were used to capture the academic, motivational, and social composition of the classrooms, respectively.

Surveys on grading practices suggest that teachers emphasize academic achievement as a grading factor more in classes with high achievement level. Research on teachers' judgment accuracy indicates that academic achievement is less accurately judged in classes characterized, for instance, by a high percentage of ethnic minorities (Martínez et al., 2009). Based on these results, we expected the test scores to be more predictive of grades in classes with a favorable student composition (high class-average test performance, interest, and parental education). Surveys on grading practices and studies analyzing actual grades suggest that noncognitive student outcomes like effort and participation are given more weight in "difficult" classes. Consequently, we assumed the interest–grade and effort–grade associations to be stronger in classes with an unfavorable student composition (low class-average test performance, interest, and parental education). Fi-

nally, research on teachers' perceptual bias indicates that grades are more biased in favor of students with a high-expectancy family background in classes with an unfavorable composition. In contrast, some studies suggest that low-expectancy students might receive a grade bonus in such classes to compensate for social disadvantages. Given these opposite predictions, no specific hypothesis was formulated on how classroom composition is related to the parental education–grade relationship.

The second research question addressed the role of teachers' classroom management. Classes with an unfavorable student composition tend to produce interruptions and disciplinary problems. Effective classroom managers should be able to establish a functioning learning environment even in these highly demanding classes. Less skilled classroom managers should be constantly engaged in cognitively demanding management activities and have less opportunity to monitor students' learning progress, providing them with a weaker basis to accurately and unbiasedly assess student performance. Consequently, grades should tend to be (a) less closely related to measures of academic knowledge and (b) more closely related to student background variables in classrooms where an unfavorable student composition and weak classroom management coincide. Thus, we expected interactions between classroom composition and classroom management when predicting the relationship between test scores and grades and parental education and grades, implying a weaker relationship between test scores and grades and a stronger relationship between parental education and grades in classes with both an unfavorable student composition and classroom management problems.

Furthermore, since ineffective classroom managers have problems in establishing a functioning learning environment, they should show increased tendency to base their grades on noncognitive outcomes like effort to ensure cooperation in classes with an unfavorable composition. Good classroom managers should less depend on grades as a means to establish order in these classes. Thus, we expected interactions between classroom composition and classroom management when predicting the relationship between noncognitive outcomes and grades, implying a stronger relationship between interest and grades, and effort and grades, in classes with both an unfavorable student composition and ineffective classroom management.

Method

Sample

This study was based on data from the MARKUS project (Helmke & Jäger, 2002), which aimed at testing the math proficiency of all eighth-grade students attending a regular secondary school type in the German federal state of Rhineland-Palatinate. The assessment was supplemented by surveys among students, math teachers, and principals of the participating schools. Tests and questionnaires were administered in May 2000 by more than 1,700 trained supervisors. Our sample included 31,038 eighth-grade students from 1,470 classes in three secondary school types (*Hauptschule* [lower school], *Realschule* [intermediate school], *Gymnasium* [highest school]). Class sizes ranged from six to 32 students ($M = 21.1$, $SD = 5.4$). The student sample was 49.5% female, and the average age was 14.8 years ($SD = 0.73$). Of the students, 14.3% reported that they had not been born in Germany. Of the students' teachers, 33.7% were female; 7.7% were under 30, 13.5% were between ages 30 and 40, 35.0% were between ages 41 and 50, 38.3% were between ages 51 and 60, and 5.5% were above 60 years old. Their average experience in teaching math was 19.2 years ($SD = 11.2$).

Variables

The study variables included the outcome variable and predictors at the student and class level. Descriptive statistics for all study variables are presented in the Appendix.

Outcome variable. Grades used in German secondary school consist of six categories from 1 to 6, with 1 representing *very good* and 6 representing *very poor* performance. The outcome variable was the student-reported math grade assigned in the report cards at the end of the first term of eighth grade. The grade was recoded so that the highest value represented the *best* performance. In the analyses, it was treated as a continuous variable, which seemed admissible as it was reasonably normally distributed (skewness = 0.088, kurtosis = −0.379).

Student-level predictors. The student-level predictors included math test performance, interest and effort in math, and parental education.

Math test performance. Students' math achievement was measured by a standardized test, developed with the aim of curricular validity by the MARKUS research group in consultation with math experts. Starting with an extensive item pool, 73 items were chosen for the final test version and distributed over eight test booklets (multimatrix design). As school types in Germany differ notably with respect to curricula and achievement level, two booklets were compiled for school types *Realschule* and *Gymnasium*, respectively, and four booklets were compiled for school type *Hauptschule*, two for each of two tracks (*Grundkurs* [basic course]; *Aufbaukurs* [advanced course]). Each booklet consisted of 15 to 17 items developed for the MARKUS test. Over and above, 15 items from the Third International Mathematics and Science Study (TIMSS) were added to each booklet, to allow for locating the student proficiencies on an international comparative scale. Preliminary analyses showed that the test could be satisfactorily described by a unidimensional Rasch model (Bond & Fox, 2007). Both TIMSS and MARKUS items were included in the scaling. Weighted-likelihood estimates (WLEs) were obtained as student proficiency measures. WLE separation reliability was 0.83.

Interest in math. Students' interest in math was assessed by the items "Working on a math problem is fun to me," "In my free time I sometimes devote myself to math over and above doing homework," "While working on a math problem, I sometimes do not notice time passing by," and "How much do you like math?" The first three items were rated on a 4-point scale and the fourth item was rated on a 5-point scale. Confirmatory factor analysis (CFA) indicated good fit of a single-factor model (comparative fit index [CFI] = 0.993, Tucker Lewis index [TLI] = 0.980, root-mean-square error of approximation [RMSEA] = 0.050, standardized root-mean-square residual [SRMR] = 0.013).[2] The items'

[2] Model fit was evaluated based on fit indices rather than the χ^2 goodness-of-fit statistic, which is highly sensitive to sample size (Bentler & Bonett, 1980). Based on recommendations in the literature (e.g., Hu & Bentler, 1999), CFI and TLI values > .90 and > .95 and RMSEA values ≤ .05 and ≤ .08 were considered as indicating acceptable and good model fit, respectively. SRMR values ≤ .08 were taken to reflect reasonable model fit.

standardized loading ranged from .53 to .82. Factor scores were obtained for the subsequent analyses. Factor score determinacy was .89, indicating sufficient factor score validity (Gorsuch, 1983). The internal consistency (Cronbach's α) of the items was .75.

Effort in math. Students' effort in math was assessed by the item "How much effort do you make in mathematics?" The item had to be rated on a 4-point scale with categories ranging from *none at all* (1) to *a lot* (4).

Parental education. Students reported their parents' graduation level on a 5-point scale from *none* to *university level*, supplemented by a category *I don't know*. In consideration of the German educational system and the international comparability of graduation levels, both variables were dichotomized into *no A-levels* and *A-levels* and recoded into two dummy variables (*one parent with A-levels* and *both parents with A-levels*, with *no parent with A-levels* as baseline category). In studying between-class variation in the parental education–grade relationship, we focused on *both parents with A-levels* (in the following denoted as "high parental education"), while *one parent with A-levels* was only used as a covariate.[3]

Class-level predictors. The class-level predictors included three classroom composition variables, classroom management, and school type.

Classroom composition. Academic, motivational, and social classroom composition were captured by the class-average test performance, interest, and parental education (the variable representing *both parents with A-levels*). While interpretation of average test performance and interest is straightforward, average parental education represented the proportion of students in a class whose parents both had A-levels (0 = 0%, 1 = 100%).

Classroom management. Previous research indicates that student ratings at the class level are valid indicators of instructional features, including classroom management (Kunter & Baumert, 2006). Four student questionnaire items with a 4-point scale were used. Two items covered the clear rules aspect ("In math lessons everyone knows the rules we are to follow," "In math lessons the teacher has clarified the consequences of rule violations"), and two items the monitoring aspect ("My math teacher always knows exactly what is happening in the classroom," "During math lessons the students are attentive and concentrated") of classroom management. In a multilevel CFA (B. O. Muthén, 1994), a model with one factor at both student and class level showed good fit to the data (CFI = 0.967, TLI = 0.900, RMSEA = 0.051, SRMR [student level] = 0.016, SRMR [class level] = 0.055). Items' class-level standardized loading ranged from .75 to 1.00.[4] The reliability of a group mean estimated from multiple assessments of a group-level property can be assessed by the ICC(2) (Bliese, 2000). Items' ICC(2) ranged from .73 to .88, indicating that the construct could be precisely measured based on the student statements. Class-level factor scores were estimated for the subsequent analyses. The internal consistency of the student ratings was Cronbach's α = .64, the internal consistency of the ratings aggregated to the class level Cronbach's α = .86.

School type. Unlike in countries such as the United States or United Kingdom, tracking in Germany is implemented between schools rather than within schools. Most secondary school students attend three school types with a long-lasting tradition: *Hauptschule*, *Realschule*, and *Gymnasium*. Data from these school types constituted our sample. Two dummy variables were created, one representing *Hauptschule* (lower school type) and one representing *Gymnasium* (highest school type), with *Realschule* (intermediate school type) as the baseline category.

Analyses

The method applied was multilevel regression analysis (Hox, 2010), with students (Level 1) nested in classes (Level 2) and the math grade as the outcome. All analyses were done in Mplus (Version 6.1; L. K. Muthén & Muthén, 2010). Initially, we specified four baseline models. Starting with the intercept-only model for the grade, test performance, interest, effort, and parental education were jointly entered as student-level predictors. Random slopes were added to represent between-class variation in the relationships between these predictors and the grade. Four models were estimated, as only one slope at a time was specified as random, while the other slopes were treated as fixed.[5] These models served as starting point for specifying cross-level interactions between student- and class-level predictors, to explain the between-class variation in the relationships between student attributes and grades.[6]

In the first analysis step, related to the first research question, the classroom composition and school type variables were entered jointly into each of the four baseline models, to explain the between-class variation in one student-level predictor's association with grades, respectively. Since our hypotheses rather unspecifically assumed an association between classroom composition and the student attribute–grade relationships, a significant effect of any of the three composition variables might have been regarded as corroborating the assumption, for example, of an effect of classroom composition on the test score–grade relationship. To control the overall probability of a Type I error, we used a Bonferroni correction on the significance tests for any set of classroom composition variables predicting a student attribute–grade relationship. In the second analysis step, related to the second research question, classroom management was entered as a class-level predictor. To examine if associations between the composition variables and the student attribute–grade relationships depended on classroom management, interaction terms between classroom

[3] Preliminary analyses had shown that (a) the overall relationship between this variable and the grade was very small and (b) the between-class variation in the relationship between this variable and the grade was noticeably smaller than for the *both parents with A-levels* variable. Thus, we decided to use the *one parent with A-levels* variable as a covariate only.

[4] A small negative residual variance resulted for one item and was fixed to zero.

[5] Preliminary analyses showed that specifying random slopes for all predictors simultaneously or specifying one for each predictor separately did not lead to any mentionable differences with respect to the random slope variance components and the associated significance tests.

[6] One problem often encountered when studying cross-level interactions is low statistical power. We used the Monte Carlo program provided by Mathieu, Aguinis, Culpepper, and Chen (2012) for a post hoc power analysis, which indicated that power in our study was high (>.9) except for very small interactions (i.e., interactions notably smaller in size than the ones considered by Mathieu et al., 2012). For each parameter identified by Mathieu et al. (2012, cf. Table 3) as influencing power, a range of values was studied. Specifically, several small values for cross-level interactions were specified as effect sizes were expected to be at best moderate. The model underlying Mathieu et al.'s power tool is not completely similar but is reasonably close to the models in this study. Among others, the tool does not consider covariates at Level 1 or Level 2.

management and composition variables were added as cross-level predictors to each of the models from Step 1.

To improve the interpretability of the regression coefficients, the outcome and all continuous predictors were standardized before entering them into the models ($M = 0$, $SD = 1$). Student-level predictors (test performance, interest, effort) were standardized based on their overall mean and variance in the sample. Class-level predictors (classroom management, classroom composition variables) were standardized based on their mean and variance across the classes (i.e., at class level). All categorical predictors (Level 1: parental education dummy variables; Level 2: school type dummy variables) were retained in their original metric.

All student-level variables except test scores contained missing data. Since traditional approaches (e.g., listwise deletion) may produce biased results, we applied multiple imputation, a state-of-the-art method to handle missing data (Schafer & Graham, 2002). In multiple imputation, each missing value is replaced by a set of predicted values, with random noise added to retain a proper amount of variability. As multiple imputation can lead to biased results if the hierarchical data structure is not taken into account, we applied the R package MICE (Van Buuren & Groothuis-Oudshoorn, 2011), which enables imputation of multivariate incomplete multilevel data based on Fully Conditional Specification (FCS). Following Schafer and Graham (2002), imputation of five values is sufficient given moderate amounts of missing data. Thus, we produced five data sets with missing data replaced by estimated values.

Results

Preliminary Analyses

Initially, we probed whether there was actual between-class variation in the student attribute–grade relationships. To this end, we specified the intercept-only model for the math grade and entered test performance, interest, effort, and both parental education dummy variables jointly as predictors at the student level. This model was estimated four times, and each time a single random slope was specified for one of four predictors (test performance, interest, effort, high parental education), while the other predictors' effects were treated as fixed. One-sided likelihood ratio tests (Hox, 2010) indicated significant variation between classes in all student attribute–grade relationships ($p < .001$, respectively).

Predicting Student Attribute–Grade Relationships by Classroom Composition

Into each of the four models described above, the classroom composition and school type dummy variables were entered jointly to predict the between-class differences in the student attribute–grade relationships. Results are shown in Table 1. Each column refers to one student-level predictor, whose relationship with the math grade was predicted by the classroom composition variables and school type. In each of these models, test performance, interest, and effort positively predicted the math grade at the student level. Better parental education was associated with a better math grade, in particular if both parents had reached A-levels. At the

Table 1
Predicting the Student Attribute–Grade Relationships by Classroom Composition Variables

Predictors	Test score		Interest		Effort		High PE	
	B	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*
Level 1: Students								
Test score	0.492***[a]	0.018	0.430***	0.009	0.433***	0.009	0.433***	0.009
Interest	0.257***	0.006	0.300***[a]	0.011	0.258***	0.006	0.260***	0.006
Effort	0.025***	0.006	0.027***	0.006	0.029*[a]	0.012	0.024***	0.006
Intermediate PE	0.038*	0.016	0.038*	0.016	0.038*	0.016	0.040*	0.016
High PE	0.176***	0.015	0.177***	0.015	0.176***	0.015	0.100**[a]	0.035
Level 2: Classes								
Grading level (intercept)								
Lower school type	0.048	0.034	0.098**	0.032	0.089**	0.032	0.085*	0.033
Highest school type	−0.050	0.033	0.002	0.031	0.002	0.031	−0.017	0.034
Average test score	−0.158***	0.021	−0.175***	0.020	−0.178***	0.020	−0.171***	0.021
Average interest	0.036**	0.011	0.041***	0.011	0.036**	0.011	0.038***	0.011
Average PE	0.033**	0.012	0.032**	0.012	0.033**	0.012	0.028*	0.013
Attribute–grade relationship (slope)								
Lower school type	−0.208***	0.030	−0.120***	0.019	−0.013	0.020	0.059	0.061
Highest school type	0.061*	0.029	0.008	0.021	0.011	0.020	0.155**	0.059
Average test score[b]	−0.052***	0.014	−0.019	0.011	−0.029*	0.010	−0.065	0.032
Average interest[b]	−0.001	0.008	−0.014	0.006	0.004	0.006	−0.012	0.018
Average PE[b]	−0.005	0.009	−0.003	0.007	0.000	0.007	0.020	0.021
ΔR^2 (slope)	8.0		6.0		5.1		—[c]	

Note. Alpha adjustment (Bonferroni) applied columnwise to class composition variables' interactions with student attributes. Intermediate PE = one parent with A-levels; High PE = both parents with A-levels; Average PE = proportion of both parents with A-levels; PE = parental education; ΔR^2 = slope variance explained over and above school type.
[a] Predicted effect when all class-level predictors equal zero. [b] Reported significance levels based on Bonferroni adjusted *p*-values. [c] Negative R^2 obtained (uninterpretable; Hox, 2010).
* $p < .05$. ** $p < .01$. *** $p < .001$.

class level, higher average test performance was related negatively and higher average interest and parental education were related positively to the classes' grading level. Slightly better grades were assigned in the lower (vs. intermediate) school type.

Results on cross-level interactions between class-level predictors and student attributes are shown in the lower part of Table 1. Since the classroom composition variables were restandardized at the class level, the intercepts refer to the student–attribute grade relation in a class from the intermediate school type with an average level of test performance, interest, and proportion of highly educated parents. In the lower school type (vs. intermediate school type), the relationships between test performance and grade and interest and grade tended to be relatively weaker. In the highest school type, the relationships between test performance and grade and high parental education and grade tended to be relatively stronger.

Contrary to our expectations, average test performance did not positively, but did *negatively*, predict the test score–grade relationship. To probe possible reasons for this finding, we removed school type from the model and estimated the cross-level effect of average test performance separately for each school type. The association of average test performance with the test score–grade relationship changed by trend from positive in the lower school type ($B = 0.030$) to negative in the highest school type ($B = -0.031$). Since at the same time the school types differed substantially in their average test performance, a *nonlinear* effect of average test performance on the test performance slope was considered. Returning to the initial model (as shown in Table 1), we removed school type but added a quadratic term for average test performance to predict the test score–grade relationship. While the linear effect of average test performance was now positive ($B = 0.074$, $SE = 0.010$, $p < .001$), the quadratic term negatively predicted the test performance slope ($B = -0.025$, $SE = 0.006$, $p < .001$). As result of this model, the relationship between test performance and grade increased as expected with an increase in class-average test performance, but only in the lower part of the average test performance spectrum. About one standard deviation above the mean of class-average test performances, a saddle point was reached, and the test score–grade relationship remained relatively stable up to the upper limit of observed class-average test performances.

We expected negative associations between the classroom composition variables and the interest–grade and effort–grade relationships. While none of the classroom composition variables predicted the interest–grade relationship, average test performance negatively predicted the association between student effort and grade ($B = -0.029$; cf. Table 1). Finally, we expected the classroom composition variables to be associated with the relationship between high parental education and grade. However, none of the classroom composition variables was significantly related to the parental education slope.

Predicting Student Attribute–Grade Relationships by Classroom Composition and Management

We expected classroom management to moderate the associations between classroom composition and the student attribute–grade relationships. We decided to confine our analysis to the interaction between classroom management and average test performance, since the latter was the only relevant predictor in the models presented so far. Classroom management, average test performance, the interaction term of both, and the school type variables were entered as cross-level predictors into each of the four random slope models introduced above. Results are shown in Table 2. Each column again refers to one student-level predictor, whose association with the grade was predicted. The student-level coefficients were comparable to the results already reported (cf. Table 1). At the class level, the interaction between average test performance and classroom management was negatively related to classes' grading level.

Results on cross-level interactions between class-level predictors and student attributes are shown in the lower part of Table 2. The school type coefficients were comparable to the results reported above (cf. Table 1). Classroom management positively predicted the test score–grade relationship ($B = 0.022$; cf. Table 2), that is, test performance tended to be more strongly related to the grade in effectively managed classes. Finally, we found one expected interaction between classroom management and average test performance, namely, when predicting the high parental education–grade relationship ($B = 0.037$). In Figure 1, the predicted high parental education coefficient is plotted against the average test performance of the class and teachers' classroom management. In line with our expectations, there was an increased predicted grade advantage of students with highly educated parents when a low average test performance of the class coincided with an ineffective classroom management.

Discussion

Summary of Findings

The present study had two research aims. First, we examined the power of academic, motivational, and social classroom composition (as captured by class-average math test performance, interest in math, and parental education) to predict the relationships of four student attributes (math test performance, interest and effort in math, high parental education) and math grades. Second, we examined the role of classroom management as a moderator of the associations between classroom composition and the student attribute–grade relationships.

Concerning the test score–grade relationship, our analysis suggested a nonlinear association with class-average test performance, as indicated by a significant quadratic term for this predictor. In line with our assumptions, the test score–grade relationship increased with average test performance, but only among classes in the lower part of the average test performance spectrum. The three classroom composition variables, supplemented by the quadratic term for average test performance, jointly explained 32.1% of the slope variance. The resulting differences in the test score–grade relationship are considerable. In classes with low average achievement (say, 2 *SD* below mean), the predicted grade difference between two students differing by one standard deviation in test performance was 0.21 *SD*; in classes with high average achievement (2 *SD* above mean) it was 0.51 *SD*.

Furthermore, average test performance negatively predicted the effort, though not the interest, relationship to grades. The classroom composition variables jointly explained 5.1% of the variation in the effort–grade relationship over and above school type. A student reporting one standard deviation higher effort than another

Table 2
Predicting the Student Attribute–Grade Relationships by Average Test Performance of the Class and Its Interaction With Classroom Management

Predictors	Test score		Interest		Effort		High PE	
	B	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*
Level 1: Students								
Test score	0.499***[a]	0.017	0.429***	0.009	0.431***	0.009	0.432***	0.009
Interest	0.260***	0.006	0.309***[a]	0.011	0.261***	0.006	0.262***	0.006
Effort	0.026***	0.006	0.027***	0.006	0.024*[a]	0.011	0.024***	0.006
Intermediate PE	0.039*	0.016	0.038*	0.016	0.039*	0.016	0.040*	0.016
High PE	0.180***	0.015	0.180***	0.015	0.180***	0.015	0.095**[a]	0.036
Level 2: Classes								
Grading level (intercept)								
Lower school type	0.074*	0.034	0.133***	0.031	0.119***	0.032	0.113**	0.033
Highest school type	−0.027	0.031	0.023	0.027	0.027	0.028	0.001	0.030
Average test score	−0.142***	0.021	−0.156***	0.020	−0.163***	0.020	−0.156***	0.021
CM	0.012	0.011	0.009	0.010	0.008	0.010	0.012	0.011
Average test score × CM	−0.030*	0.012	−0.025*	0.010	−0.019	0.010	−0.026*	0.011
Attribute–grade relationship (slope)								
Lower school type	−0.231***	0.029	−0.135***	0.019	0.004	0.020	0.090	0.066
Highest school type	0.064*	0.026	0.003	0.018	0.004	0.018	0.177**	0.057
Average test score	−0.059***	0.014	−0.022	0.011	−0.022*	0.011	−0.054	0.031
CM	0.022*	0.010	−0.003	0.006	−0.009	0.006	−0.035	0.021
Average test score × CM	−0.007	0.007	−0.005	0.006	0.004	0.006	0.037*	0.017
ΔR^2 (slope)	11.0		3.6		5.9		—[b]	

Note. Intermediate PE = one parent with A-levels; High PE = both parents with A-levels; PE = parental education; CM = classroom management; ΔR^2 = slope variance explained over and above school type.
[a] Predicted effect when all class-level predictors equal zero. [b] Negative R^2 obtained (uninterpretable; Hox, 2010).
* $p < .05$. ** $p < .01$. *** $p < .001$.

was predicted to have a 0.09 *SD* grade advantage in classes with low average achievement (2 *SD* below mean), and a slight 0.03 *SD* disadvantage in classes with high average achievement (2 *SD* above mean). This is a small but not irrelevant effect, since large effort differences within classes should be common.

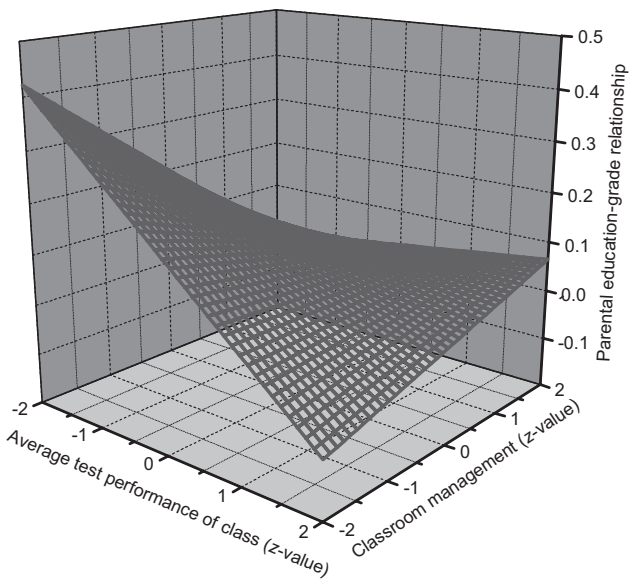

Figure 1. Predicted effect of average test performance of the class and teachers' classroom management on the relationship between high parental education and math grade.

Finally, we found evidence for an interaction between average test performance and classroom management in predicting the relationship between parental education and grades. As expected, high parental education was most predictive of grades when ineffective classroom managers were confronted with a low-performing class. If both average test performance and classroom management were two standard deviations below their mean, the predicted grade advantage of students with highly educated parents was 0.42 *SD*. In contrast, if average test performance was low (2 *SD* below mean) but classroom management was excellent (2 *SD* above mean), this predicted grade advantage was close to zero ($B = -0.02$).

Classroom Composition and Teachers' Grading Practices

Our results are in line with teachers' self-reported practice to put more emphasis on achievement and less emphasis on nonachievement factors when determining grades in high ability classes (McMillan, 2001). They are also consistent with studies indicating that noncognitive outcomes are more strongly related to grades in "difficult" classes or schools (Howley et al., 2000; Rakoczy et al., 2008). Substantively, our findings are compatible with a view on grades as a tool of instruction that is used by teachers in order to serve important goals of instruction but also as a means to establish control. In particular, the emphasis on noncognitive outcomes in demanding classes may serve both to encourage students to engage in learning (Brookhart, 1993; McMillan, 2001) and to ensure a basic level of cooperation (Ennis, 1995). From a larger perspective, such grading practices might be considered as one aspect of

various strategies that teachers apply to adapt their instruction to students' learning prerequisites. In particular, average achievement of the class—the only composition variable of predictive value in this study—is relevant to teachers' decisions on the organizational, curricular, and didactical features of instruction (Dreeben & Barr, 1988) and has been found related to several aspects of instructional quality (Weinert, Schrader, & Helmke, 1989).

The correlation of test scores and grades has traditionally been a main criterion for the validity of grades (Bowers, 2009). Moreover, measurement experts have advised teachers not to consider noncognitive outcomes in grading (Cross & Frary, 1999; Frary, Cross, & Weber, 1993). Notwithstanding, the predicted test score–grade relation approached zero in low-performing classes, while effort was more strongly related to grades. Should we thus worry about the grading practices in these classes? Substantial arguments have been brought forward against the psychometrically oriented viewpoint described above (Bowers, 2009; Willingham et al., 2002). Nevertheless, we would argue that our results might indicate a somewhat higher prevalence of debatable grading practices in low-performing classes. For example, some teachers might tend to largely abandon the curriculum in order to avoid confrontations and assign grades based mostly on students' willingness to cooperate (Ennis, 1995, 1996).

Classroom Composition and Accuracy/Bias in Teachers' Grading

We expected that students' academic knowledge should be more accurately judged in classes with a favorable student composition, enhancing the relation between test scores and grades. The finding that average test performance is positively related to the test score–grade association among lower-performing classes gives some support to this claim. The assumed explanation is that in "difficult" classes teachers are more involved in management activities and less involved in achievement-related interactions, limiting their possibility to form accurate judgments. However, if less attentiveness and more behavioral problems impair teachers' judgments, it seems surprising that average achievement but not interest has been found to predict the test score–grade relationship. Then again, this relationship was found to be stronger in effectively managed classrooms (cf. Table 2). This constellation might suggest that it is not primarily students' classroom behavior but rather teachers' lack of management skills that reduces opportunities to observe student performance and thus diminishes accuracy.

While there was no global association between classroom composition variables and the parental education–grade relationship, average test performance of the class and classroom management were found to significantly interact in predicting this relationship. In this situation, a noticeable grade advantage of students with high parental education was found, potentially indicating bias against low-SES students. This result is in line with our theoretical considerations on perceptual bias. We had hypothesized a grade advantage in favor of high-expectancy students to be related to a "difficult" classroom environment that is inefficiently organized. On the other hand, we had speculated that teachers might feel a need in these classes to compensate for social disadvantages and might adjust grades in favor of disadvantaged students, but our results contradict this latter notion.

Educational Implications

Given the cross-sectional nature of data, it is important not to overinterpret the results. Notwithstanding, we point to two possible implications concerning teacher training. First, since classroom composition is typically difficult to change, classroom management would seem the most promising variable in this study to focus on. Improving classroom management might have some desirable effects on teachers' grading, as it was found to be negatively related to the parental education effect in low-performing classes and positively related to the relationship between test scores and grades (provided the latter should be raised, see above).

Furthermore, some studies have indicated that assessment practices of teachers can be improved by measurement training (e.g., Bonner & Chen, 2009). It might prove useful to enhance teachers' awareness of potential problems in grading related to the classroom context and their interaction with students. For instance, trainers might point out that evaluating students can be more tedious and that even more carefulness is required in classrooms with an unfavorable student composition, since judgments may be more often inaccurate and biased.

Limitations

Several limitations derive from our use of student reports. Student self-reported grades were analyzed in this study. Among others, the validity of self-reported grades is moderated by student performance, with grades being less validly reported by low-performing students (Kuncel, Credé, & Thomas, 2005). This may have contributed to at least one finding, namely, the lower test score–grade association in low-performing classes. Other student-reported information can be incorrect, too. For instance, parental education is less reliably reported by low-performing students, which may lead to bias in estimates of regression coefficients (Kreuter, Eckman, Maaz, & Watermann, 2010). Furthermore, our assessment of student attributes and classroom management was based on few items exclusively from student questionnaires. More extensive questionnaires and additional sources of information (e.g., external observers) would allow for a more profound assessment of these variables.

The generalizability of our findings is limited in several respects. First, we only analyzed grades assigned by math teachers. Teachers in subjects like English or social studies seem to be more flexible in their consideration of achievement and nonachievement factors (Frary et al., 1993; McMillan, 2001) and thus might have their grading practices even more affected by classroom composition or classroom management issues. Second, our sample included only secondary school classes. Teachers in secondary school consider student misbehavior more than elementary school teachers when assigning grades (Randall & Engelhard, 2009), which might result from increased challenges in classroom management. If so, the relation between classroom composition and the use of variables such as effort in determining grades might be less pronounced in elementary school. Third, cultural differences might affect the results. Teachers' classroom management and students' classroom discipline were found to vary between cultures to some extent (Lewis, Romi, Qui, & Katz, 2005). This might imply that certain grading strategies play a larger role in some cultures than in others.

To conclude, the present study highlights the significance of classroom composition for the relationship between different types of student attributes and grades. Our findings suggest that teachers adapt their grading practices to classroom conditions by laying more emphasis on academic achievement or nonachievement factors but also that making judgments that accurately and unbiasedly reflect student outcomes is more challenging in "difficult" classes. To some extent, effective classroom management appears to be a protective factor against potentially problematic aspects of grading, and enhancing classroom management skills might be one way to enhance quality of teachers' evaluation of students. A task for future research will be to study in detail the processes mediating between classroom composition, classroom management, and the student attribute relationships with grades.

References

Barth, J. M., Dunlap, S. T., Dane, H., Lochman, J. E., & Wells, K. C. (2004). Classroom environment influences on aggression, peer relations, and academic focus. *Journal of School Psychology, 42,* 115–133. doi:10.1016/j.jsp.2003.11.004

Bennett, R. E., Gottesman, R. L., Rock, D. A., & Cerullo, F. (1993). Influence of behavior perceptions and gender on teachers' judgments of students' academic skill. *Journal of Educational Psychology, 85,* 347–356. doi:10.1037/0022-0663.85.2.347

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588–606. doi:10.1037/0033-2909.88.3.588

Blair, I. V., & Banaji, M. R. (1996). Automatic and controlled processes in stereotype priming. *Journal of Personality and Social Psychology, 70,* 1142–1163. doi:10.1037/0022-3514.70.6.1142

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In S. W. J. Kozlowski & K. J. Klein (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 349–381). San Francisco, CA: Jossey-Bass.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model*. Mahwah, NJ: Erlbaum.

Bonner, S. M., & Chen, P. P. (2009). Teacher candidates' perceptions about grading and constructivist teaching. *Educational Assessment, 14,* 57–77. doi:10.1080/10627190903039411

Bowers, A. J. (2009). Reconsidering grades as data for decision making: More than just academic knowledge. *Journal of Educational Administration, 47,* 609–629. doi:10.1108/09578230910981080

Bowers, A. J. (2011). What's in a grade? The multidimensional nature of what teacher-assigned grades assess in high school. *Educational Research and Evaluation, 17,* 141–159. doi:10.1080/13803611.2011.597112

Brookhart, S. M. (1993). Teachers' grading practices: Meaning and values. *Journal of Educational Measurement, 30,* 123–142. doi:10.1111/j.1745-3984.1993.tb01070.x

Brookhart, S. M. (1994). Teachers' grading: Practice and theory. *Applied Measurement in Education, 7,* 279–301. doi:10.1207/s15324818ame0704_2

Brookhart, S. M. (2004). Classroom assessment: Tensions and intersections in theory and practice. *Teachers College Record, 106,* 429–458. doi:10.1111/j.1467-9620.2004.00346.x

Cothran, D. J., & Ennis, C. D. (1997). Students' and teachers' perceptions of conflict and power. *Teaching and Teacher Education, 13,* 541–553. doi:10.1016/S0742-051X(97)85542-4

Cross, L. H., & Frary, R. B. (1999). Hodgepodge grading: Endorsed by students and teachers alike. *Applied Measurement in Education, 12,* 53–72. doi:10.1207/s15324818ame1201_4

Dreeben, R., & Barr, R. (1988). Classroom composition and the design of instruction. *Sociology of Education, 61,* 129–142. Retrieved from http://www.jstor.org/stable/2112622. doi:10.2307/2112622

Dusek, J. B., & Joseph, G. (1983). The bases of teacher expectancies: A meta-analysis. *Journal of Educational Psychology, 75,* 327–346. doi:10.1037/0022-0663.75.3.327

Emmer, E. T., & Stough, L. M. (2001). Classroom management: A critical part of educational psychology, with implications for teacher education. *Educational Psychologist, 36,* 103–112. doi:10.1207/S15326985EP3602_5

Ennis, C. D. (1995). Teachers' responses to noncompliant students: The realities and consequences of a negotiated curriculum. *Teaching and Teacher Education, 11,* 445–460. doi:10.1016/0742-051X(95)00010-H

Ennis, C. D. (1996). When avoiding confrontation leads to avoiding content: Disruptive students' impact on curriculum. *Journal of Curriculum and Supervision, 11,* 145–162. Retrieved from http://www.ascd.org/publications/jcs/winter1996/toc.aspx

Evertson, C. M. (1982). Differences in instructional activities in higher- and lower-achieving junior high English and math classes. *The Elementary School Journal, 82,* 329–350. doi:10.1086/461271

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 23, pp. 1–74). New York, NY: Academic Press. doi:10.1016/S0065-2601(08)60317-2

Frary, R. B., Cross, L. H., & Weber, L. J. (1993). Testing and grading practices and opinions of secondary teachers of academic subjects: Implications for instruction in measurement. *Educational Measurement: Issues and Practice, 12,* 23–30. doi:10.1111/j.1745-3992.1993.tb00539.x

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review, 102,* 652–670. doi:10.1037/0033-295X.102.4.652

Gilbert, D. T., & Osborne, R. E. (1989). Thinking backward: Some curable and incurable consequences of cognitive busyness. *Journal of Personality and Social Psychology, 57,* 940–949. doi:10.1037/0022-3514.57.6.940

Gorsuch, R. L. (1983). *Factor analysis*. Hillsdale, NJ: Erlbaum.

Hallinan, M. T. (1992). The organization of students for instruction in the middle school. *Sociology of Education, 65,* 114–127. doi:10.2307/2112678

Helmke, A., & Jäger, R. S. (Eds.). (2002). *Das Projekt MARKUS. Mathematik-Gesamterhebung Rheinland-Pfalz: Kompetenzen, Unterrichtsmerkmale, Schulkontext* [The MARKUS Project: Mathematics survey in the German federal state of Rhineland-Palatinate: Competencies, features of instruction, school context]. Landau, Germany: Verlag Empirische Pädagogik.

Howley, A., Kusimo, P. S., & Parrott, L. (2000). Grading and the ethos of effort. *Learning Environments Research, 3,* 229–246. doi:10.1023/A:1011469327430

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications*. New York, NY: Routledge.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55. doi:10.1080/10705519909540118

Jussim, L. (1989). Teacher expectations: Self-fulfilling prophecies, perceptual biases, and accuracy. *Journal of Personality and Social Psychology, 57,* 469–480. doi:10.1037/0022-3514.57.3.469

Klapp Lekholm, A. (2011). Effects of school characteristics on grades in compulsory school. *Scandinavian Journal of Educational Research, 55,* 587–608. doi:10.1080/00313831.2011.555923

Klapp Lekholm, A., & Cliffordson, C. (2008). Discrepancies between school grades and test scores at individual and school level: Effects of gender and family background. *Educational Research and Evaluation, 14,* 181–199. doi:10.1080/13803610801956663

Kounin, J. S. (1970). *Discipline and group management in classrooms*. New York, NY: Holt, Rinehart and Winston.

Kreuter, F., Eckman, S., Maaz, K., & Watermann, R. (2010). Children's reports of parents' education level: Does it matter whom you ask and what you ask about? *Survey Research Methods, 4,* 127–138. Retrieved from https://ojs.ub.uni-konstanz.de/srm/

Kruglanski, A. W., & Freund, T. (1983). The freezing and unfreezing of lay-inferences: Effects on impressional primacy, ethnic stereotyping, and numerical anchoring. *Journal of Experimental Social Psychology, 19,* 448–468. doi:10.1016/0022-1031(83)90022-7

Kruglanski, A. W., Pierro, A., Mannetti, L., Erb, H.-P., & Chun, W. Y. (2007). On the parameters of human judgment. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 39, pp. 255–303). New York, NY: Academic Press.

Kuncel, N. R., Credé, M., & Thomas, L. L. (2005). The validity of self-reported grade point averages, class ranks, and test scores: A meta-analysis and review of the literature. *Review of Educational Research, 75,* 63–82. doi:10.3102/00346543075001063

Kunter, M., & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research, 9,* 231–251. doi:10.1007/s10984-006-9015-7

Kunter, M., Baumert, J., & Köller, O. (2007). Effective classroom management and the development of subject-related interest. *Learning and Instruction, 17,* 494–509. doi:10.1016/j.learninstruc.2007.09.002

LePage, P., Darling-Hammond, L., Akar, H., Gutierrez, C., Jenkins-Gunn, E., & Rosebrock, K. (2005). Classroom management. In L. Darling-Hammond & J. Bransford (Eds.), *Preparing teachers for a changing world* (pp. 327–357). San Francisco, CA: Jossey-Bass.

Letzring, T. D., Wells, S. M., & Funder, D. C. (2006). Information quantity and quality affect the realistic accuracy of personality judgment. *Journal of Personality and Social Psychology, 91,* 111–123. doi:10.1037/0022-3514.91.1.111

Lewis, R., Romi, S., Qui, X., & Katz, Y. J. (2005). Teachers' classroom discipline and student misbehavior in Australia, China and Israel. *Teaching and Teacher Education, 21,* 729–741. doi:10.1016/j.tate.2005.05.008

Lysne, A. (1984). Grading of student's attainment: Purposes and functions. *Scandinavian Journal of Educational Research, 28,* 149–165. doi:10.1080/0031383840280303

Macrae, C. N., Hewstone, M., & Griffiths, R. J. (1993). Processing load and memory for stereotype-based information. *European Journal of Social Psychology, 23,* 77–87. doi:10.1002/ejsp.2420230107

Martínez, J. F., Stecher, B., & Borko, H. (2009). Classroom assessment practices, teacher judgments, and student achievement in mathematics: Evidence from the ECLS. *Educational Assessment, 14,* 78–102. doi:10.1080/10627190903039429

Mathieu, J. E., Aguinis, H., Culpepper, S. A., & Chen, G. (2012). Understanding and estimating the power to detect cross-level interaction effects in multilevel modeling. *Journal of Applied Psychology, 97,* 951–966. doi:10.1037/a0028380

McMillan, J. H. (2001). Secondary teachers' classroom assessment and grading practices. *Educational Measurement: Issues and Practice, 20,* 20–32. doi:10.1111/j.1745-3992.2001.tb00055.x

Merrett, F., & Wheldall, K. (1992). Teacher training and classroom discipline. In K. Wheldall (Ed.), *Discipline in schools* (pp. 10–19). London, England: Routledge.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods and Research, 22,* 376–398. doi:10.1177/0049124194022003006

Muthén, L. K., & Muthén, B. O. (2010). *Mplus* (Version 6.1) [Computer software]. Los Angeles, CA: Muthén & Muthén.

Opdenakker, M.-C., Van Damme, J., De Fraine, B., Van Landeghem, G., & Onghena, P. (2002). The effect of schools and classes on mathematics achievement. *School Effectiveness and School Improvement, 13,* 399–427. doi:10.1076/sesi.13.4.399.10283

Pace, J. L., & Hemmings, A. (2007). Understanding authority in classrooms: A review of theory, ideology, and research. *Review of Educational Research, 77,* 4–27. doi:10.3102/003465430298489

Pratto, F., & Bargh, J. A. (1991). Stereotyping based on apparently individuating information: Trait and global components of sex stereotypes under attention overload. *Journal of Experimental Social Psychology, 27,* 26–47. doi:10.1016/0022-1031(91)90009-U

Punch, K. F., & Tuettemann, E. (1990). Correlates of psychological distress among secondary school teachers. *British Educational Research Journal, 16,* 369–382. doi:10.1080/0141192900160405

Rakoczy, K., Klieme, E., Bürgermeister, A., & Harks, B. (2008). The interplay between student evaluation and instruction. Grading and feedback in mathematics classrooms. *Journal of Psychology, 216,* 111–124. doi:10.1027/0044-3409.216.2.111

Randall, J., & Engelhard, G. (2009). Differences between teachers' grading practices in elementary and middle schools. *The Journal of Educational Research, 102,* 175–186. doi:10.3200/JOER.102.3.175-186

Rindermann, H. (2007). Die Bedeutung der mittleren Klassenfähigkeit für das Unterrichtsgeschehen und die Entwicklung individueller Fähigkeiten [The relevance of class ability for teaching and development of individual competences]. *Unterrichtswissenschaft, 35,* 68–89. Retrieved from http://www.pedocs.de

Roland, E., & Galloway, D. (2002). Classroom influences on bullying. *Educational Research, 44,* 299–312. doi:10.1080/0013188022000031597

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7,* 147–177. doi:10.1037/1082-989X.7.2.147

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software, 45,* 1–67. Retrieved from http://www.jstatsoft.org/v45/i03

Weinert, F. E., Schrader, F.-W., & Helmke, A. (1989). Quality of instruction and achievement outcomes. *International Journal of Educational Research, 13,* 895–914. doi:10.1016/0883-0355(89)90072-4

Willingham, W. W., Pollack, J. M., & Lewis, C. (2002). Grades and test scores: Accounting for observed differences. *Journal of Educational Measurement, 39,* 1–37. doi:10.1111/j.1745-3984.2002.tb01133.x

(*Appendix follows*)

Appendix

Descriptive Statistics and Correlations, Variances, and Covariances of Study Variables

Variable	*M*	*SD*	Skewness	Kurtosis	% Missing	1	2	3	4	5	6
Student level											
1. Mathematics grade[a]	3.19	1.02	.09	−.38	5.92	**1.04**	−.34	−.14	−.08	−.01	−.03
2. Mathematics test performance (WLE)	.00	.98	−.04	−.87	0.00	−.34	**.96**	.05	−.05	.05	.06
3. Interest in mathematics (factor scores)	.00	.40	.28	−.72	4.18	−.35	.14	**.16**	.10	<−.01	<.01
4. Effort in mathematics	1.73	.77	−.16	−.35	6.29	−.10	−.07	.32	**.59**	−.01	−.01
5. Intermediate parental education	.20	.40	1.26	−.41	34.84[b]	−.03	.12	−.01	−.04	**.16**	−.03
6. High parental education	.13	.34	1.57	.45	34.84[b]	−.10	.18	.02	−.05	−.20	**.12**
Class level											
1. Average test performance	−.09	.74	−.07	−.74	0.00	**.55**	−.02	.05	−.04	−.27	.23
2. Average interest in mathematics	.01	.13	.17	.12	0.00	−.19	**.02**	<−.01	−.02	.03	−.01
3. Average parental education	.12	.13	.98	.64	0.00	.58	−.20	**.02**	−.01	−.03	.04
4. Classroom management (factor scores)	.00	.34	−.62	.27	0.00	−.15	.50	−.28	**.12**	.06	−.05
5. Lower school type (Hauptschule)	.44	.50	.25	−1.94	0.00	−.74	.43	−.53	.37	**.25**	−.13
6. Highest school type (Gymnasium)	.29	.46	.91	−1.12	0.00	.67	−.23	.76	−.30	−.57	**.21**

Note. Student level $n = 31{,}038$. Class level $n = 1{,}470$. Correlations are in the lower left triangle, variances are on the diagonal, and covariances are in the upper right triangle. All statistics reported are based on the original variables, before applying multiple imputation. WLE = Warm's weighted-likelihood estimate.

[a] Mathematics grade in its original metric with 1 = *very good*, 6 = *very poor* performance. [b] Percentage of students with information missing for either education of mother, or education of father, or both. Percentage missing for mother's education: 27.4%, and for father's education: 30.5%.

Received October 17, 2011
Revision received June 10, 2013
Accepted June 24, 2013 ■