ABSTRACT

Title of Document: ELEMENTARY TEACHERS' GRADING PRACTICES: DOES THE REALITY REFLECT THE RUBRIC?

Katherine B. Shanahan, Doctor of Philosophy, 2011

Directed By: Professor, Gary Gottfredson, Counseling and Personnel Services

Report cards are the primary way that teachers, students, and parents communicate about student achievement in the classroom. Although many school districts develop rubrics to guide teacher grading practices, most research finds that in reality, grades represent a hodgepodge of factors that vary across teachers and across school systems. The current study investigates student factors that explain variance in elementary report card grades in a suburban school district. The sample includes 4th and 5th grade students (N = 8,555) and their classroom teachers (N = 374) from 45 schools. Multilevel structural equation models, with students nested within classrooms, tested two models describing variance in report card grades. One model included the factors listed on the school system grading rubric along with additional factors thought to be related to grades (non-rubric model). An alternative, nested, model included only the factors from the grading rubric (rubric model). Results

suggested that the non-rubric model provided a better fit for the data, but effects for the additional non-rubric factors were uniformly small.

ELEMENTARY TEACHERS' GRADING PRACTICES: DOES THE REALITY REFLECT THE RUBRIC?

By

Katherine Bruckman Shanahan

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2011

Advisory Committee:
Professor Gary Gottfredson, Chair
Associate Professor Bill Strein
Professor Dennis Kivlighan
Assistant Professor Matthew Miller
Assistant Professor Jeffrey Harring
Professor Denise Gottfredson, Dean's Representative

Acknowledgements

Katherine Bruckman Shanahan, Counseling and Personnel Services, University of Maryland.

This research was supported in part by Institute of Education Sciences (IES) grant #R305F050051. Opinions expressed do not necessarily reflect those of the sponsor. I am grateful for the contributions of the IC Teams evaluation research group at the University of Maryland (Gary Gottfredson, Sylvia Rosenfield, Phuong Vu, Jill Berger, Megan Vaganek, Deborah Nelson, Todd Gravois, and Eva Yiu). I especially wish to thank Gary Gottfredson for his support, advice, and manuscript editing.

Correspondence concerning this article should be addressed to Katherine B. Shanahan, Department of Counseling and Personnel Services, University of Maryland, College Park, MD 20742. E-mail: kjb.shanahan@gmail.com.

Table of Contents

List of Tables

List of Figures

Elementary Teachers' Grading Practices: Does the Reality Reflect the Rubric?

Report cards are the primary formal ways that schools and parents communicate about student achievement in the classroom (Allen, 2005; Friedman & Frisbie, 1995). Dreaded by some students and loved by others, grades are a key part of students' schooling from kindergarten to university. Although grades are important communication tools, they are often confusing and misunderstood by students, parents, and teachers (Allen, 2005). Grading can be a complicated task for teachers because it challenges them to rate student performance in an appropriate and fair way (Carlson, 2003).

One reason that grades can be difficult to interpret is that teachers use judgment when assigning grades, and consider many factors (not just achievement) when grading students (Brookhart, 1993). Research has found that teachers use a hodgepodge of factors when grading students, but that academic achievement is the largest factor (Brookhart, 1994; Bursuck, Polloway, Plante, Epstein, Jayanthi, & McConeghy, 1996; Cross & Frary, 1996; McMillan, Myran & Workman, 2002; Strein & Meshbesher, 2006). Generally, researchers investigating teacher grading practices ask teachers to complete questionnaires designed to measure the degree to which they incorporate different factors into their grading (Bursuck et al., 1996; Cross & Frary, 1996; Frary, Cross & Weber, 1992; McMillan et al., 2002), or ask teachers to respond to hypothetical grading scenarios (Brookhart, 1993; Brookhart, 1994).

Teacher Grading Practices

A study by McMillan et al. (2002) examined grading practices in a large sample of teachers and schools, and will be discussed here in detail both because of its relevance to the current study and because it represents a current trend in research on grading practices. The researchers examined upper elementary (grades 3-5) teachers' assessment and grading practices. The sample included 901 teachers in 105 schools in seven metropolitan Virginia school districts. A teacher self-report questionnaire measured the degree to which teachers used different assessment and grading practices (McMillan et al., 2002). Teachers rated the extent to which they incorporated specific factors into their grading practices using a 6-point scale ranging from *not at all* to *completely*. Descriptive analyses summarized teacher responses, and multiple regression and paired *t* tests examined relations between variables. Teachers reported that disruptive behaviors of students contributed *very little* to their grading practices. Student academic performance and mastery of learning goals contributed *quite a bit* to *extensively* to their grading practices. Other variables, such as student effort, work habits, participation and/or attention, contributed *very little* to *quite a bit*, and had large standard deviations, indicating greater teacher variability in the use of these factors.

Several recent studies used more sophisticated methods to examine factors that contribute to grades (Lekholm & Cliffordson, 2008; Randall & Engelhard, 2009; Randall & Engelhard, 2010). Similar to McMillan et al. (2002), Randal and Engelhard (2010) also surveyed public school teachers ($N = 516$) in a metropolitan school district. Teachers read scenarios describing student ability, achievement,

behavior, and effort, and assigned grades to rate the scenarios. Although results indicated that teachers graded mostly based on achievement, they also considered non-achievement factors, especially when scenarios seemed on the borderline of a lower grade versus a higher one.

Although descriptive studies of grading practices may show teacher perceptions of grading practices, they do not analytically indicate the factors that influence student grades. Lekholm & Cliffordson (2008) addressed this gap in the grading research by using confirmatory factor analysis to identify and examine factors that contribute to grades. Participants included 99,070 ninth grade students from 1,246 schools in Sweden. Results indicated that the largest proportion of variance in grades was accounted for by achievement, as measured by national standardized tests, and also identified an additional, non-achievement dimension that explained variance in grades. The items representing the non-achievement factor were student sex and parent educational attainment. The researchers hypothesized that the non-achievement factor represented student behaviors or student characteristics.

In general, grading practices are examined by asking teachers either to describe how they grade students or to answer questionnaires about their practices. When asked to self-report their practices, teachers report that academic performance contributes the most to their grading practices, and sometimes indicate that they consider non-achievement factors as well (e.g. McMillan et al., 2002; Randall & Engelhard, 2010). Sometimes, questionnaire studies do not even ask teachers to indicate their use of non-achievement factors when assigning grades (Bursuck, et al., 1996).

Teachers may not consistently self-report that they consider non-achievement factors when grading students. Some evidence indicates that non-achievement factors account for some variance in grades, but the researchers did not structure their study to identify or describe these factors (Lekholm & Cliffordson; 2008). Other research shows that many variables, such as problem behavior, effort, and student characteristics, are related to the grades students receive (e.g. Bruckman, 2010; Duckworth & Seligman, 2006; Gottfredson, 1981; Hinshaw, 1992; Randall & Engerhald, 2010). Examination of the literature on relations between student achievement, student behaviors, and student characteristics can help to clarify the non-achievement factors that influence grades.

Influence of Student Behavior and Demographic Characteristics on Grades

Research consistently finds a relation between student problem behavior and student achievement (Bruckman, 2010; Bubb, McCartney & Willett, 2007; Crosby & French, 2002; Gottfredson, 1981; Hinshaw, 1992; Johnson, McGue & Ianoco, 2006). Prior research, however, has not measured these constructs in consistent ways. Studies often measured achievement with group-administered classroom tests, rather than with grades (Bubb et al., 2007; Crosby & French, 2002; Hinshaw, 1992) and measured problem behavior with parent ratings or delinquency records rather than with classroom behavior measures (Gottfredson, 1981; Johnson et al., 2006). Although some studies do report correlations between classroom problem behavior and grades, these results are not usually the focus of the research (e.g. Gottfredson & Gottfredson, 1999).

Some evidence suggests that teachers believe that student behavior should affect the grades they receive (Frary et al., 1992) and that they consider problem behavior when assigning grades (Randall & Engelhard, 2010). Indeed, the relation between behavior and achievement may be stronger when achievement is measured with teacher ratings (such as grades) than when measured with standardized tests (Alexander, Entwisle, & Dauber, 1993). This suggests that teachers may take behavior into account when rating student academic performance on the report card. This possibility is plausible. Other research examining the relation between student variables (ethnicity, gender, socioeconomic status, etc.), teacher ratings of achievement, and standardized test scores suggests that the relation between some student variables and grades may be stronger than their relation with performance on standardized tests (Beswick, Willms & Sloat, 2005; Martinez, Stecher & Borko, 2009; Stone, 1994).

For example, Beswick et al. (2005) investigated the discrepancy between teacher ratings and standardized measures of literacy, and examined whether other variables such as student behavior, family characteristics, or SES could explain the discrepancy. Nine Canadian schools participated in the study, including 205 kindergarteners and 12 teachers. The *Teacher Rating Scale (TRS)—Literacy* (Flynn, 1997) served as a teacher rating of literacy, and the *Word Reading* subtest from the *Wechsler Individual Achievement Test (WIAT)—Second Edition* (Psychological Corporation, 2002) provided a standardized measure. Teachers completed the Conners' Teacher Rating Scale (Conners, 2001), a rating of student behavior, for each student, and schools provided student demographic information including gender,

retention, age, and family background. Correlations between behavior scales and teacher-rated literacy were greater than those between the behavior scales and the standardized literacy scores. Researchers also calculated difference scores between the two literacy measures by subtracting standardized raw scores on the TRS from scores on the WIAT. Behavior ratings significantly predicted the discrepancy, beyond the influence of child gender, parental education, and mother's work. The researchers concluded that teacher ratings seem affected by child and family characteristics and that child gender and behavior were most influential.

Although the Beswick et al. (2006) study points to child factors, other than achievement, which influence teacher ratings of academic skills, some considerations limit causal inferences that can be made. One limitation is the reliance on discrepancy scores, which are unreliable. It would make more sense to look for predictors of the two different measures of literacy, and compare the regression coefficients, or to use hierarchical multiple regression, rather than to look for predictors of the discrepancy scores for the two literacy measures

Bruckman (2010) focused specifically on the relation between problem behavior and grades. Multilevel models, with students nested within classrooms, tested the influence of student problem behavior (as rated by teachers) on student GPA, math grades, and reading grades. Results implied that problem behavior negatively influenced grades for students at each grade level, controlling for standardized academic achievement and other student and classroom-level covariates. Results supported the idea that grades include factors other than academic

achievement; the problem behavior factor made a robust contribution to grades across subject areas and grade levels.

Student problem behavior seems to influence student grades. In addition, teachers seem to support including their ratings of student effort in the grades they assign (Allen, 2005; Frary et al., 1992), and report that effort influences their grading (Randall & Engelhard, 2010). Research also finds consistent associations between academic performance and student characteristics such as sex, socio-economic status, and ethnicity. Girls tend to receive higher grades than boys (Duckworth & Seligman, 2006; Pomerantz, Altermatt & Saxon, 2002). Students from higher socio-economic status tend to perform better than those from lower levels of SES (Hanushek & Luque, 2003). And finally, students who are African American or Hispanic tend to receive lower grades than Caucasian and Asian students (Herman, 2009). These associations of student characteristics with grades could be due to differences in student behavior. For example, girls, White and Asian students, and higher SES students may put forth greater effort in the classroom, and this increased effort could lead to increased grades. Alternatively, these students might simply be perceived as putting forth greater effort, which could reflect bias. A variety of speculative explanations can explain the associations of demographic characteristics with grades.

A large body of research, beyond the scope of the current study, examines the complex processes that underlie the relation between race and academic achievement. For example, some researchers have hypothesized that parenting practices, parental education, and family socioeconomic status could mediate the relation between race and academic achievement (e.g. Bodovski, 2010; Davis-Kean & Sexton, 2009).

Other researchers have hypothesized that teacher biases and prejudice could underlie the relation of race with achievement (e.g. Vaught, S. & Castagno, 2008).

Grading Policies

Educators, policy makers, and researchers recognize that grades often measure non-achievement factors, and that grading practices are often inconsistent (e.g. Greville, 2009; Allen, 2005). Many school systems address these grading challenges by developing policies to guide teacher practices (Polloway et al., 1994; Strein & Meshbesher, 2006). These policies tend to differ across school districts (Austin & McCann, 1992; Polloway et al., 1994). Grading rubrics, which describe how grades should correspond to different levels of performance and state the factors that should contribute to grading, are one approach for developing grading policy (Greville, 2009).

Grading policies are widespread, but not necessarily effective in guiding teacher grading practices. Research examining whether grades reflect the constructs teachers are instructed to incorporate into their grading is scarce. Some evidence indicates that teachers often do not follow the guidelines (Strein & Meshbesher, 2006). For example, grades may measure factors other than those intended by the grading policy, as suggested by Bruckman (2010). I found consistent associations across subject areas between problem behavior and grades even though the rubric which guided teacher grading did not instruct teachers to include problem behavior in their grading of students. However, limitations of the multilevel modeling approach, and omission of key variables such as student effort, threatened inferences regarding how the students' grades matched with the grading policy.

Limitations of Prior Research

Grades are viewed as representing a confusing hodgepodge of achievement and non-achievement factors. Grading rubrics, a form of grading policy, are intended to describe the factors that should account for variance in grades, yet little (if any) research has examined whether these rubrics are followed when assigning grades.

Many previous studies were also hindered by methodological limitations. Most were descriptive and relied on teacher self-reports of their grading practices, or their responses to hypothetical scenarios. Only one study examined factors that account for the variance in grades (Lekholm & Cliffordson, 2008), and none of the previous research modeled grades at both the student and classroom levels. A more sophisticated approach would be to consider both the multilevel nature of student report card grade data, and the structural relations among variables associated with grades.

Purpose

The purpose of this study is to examine the factors associated with student grades and explore whether teachers in a large suburban school district seem to follow the district grading rubric when assigning grades. According to the rubric, teachers in the district are to consider student academic achievement, effort, and attendance when assigning grades (see Appendices A and B). This study is not an attempt to develop a complete causal model of grades, but rather an attempt to describe the factors that are associated with grades by comparing a hypothesized (non-rubric) model to an alternative (rubric-only) model.

Research Questions

1. Do report card grades reflect academic achievement, effort, and attendance factors outlined on the district's grading rubric?

2. What are the relative associations of achievement, effort, and attendance with the grades students receive?

3. Does a non-rubric model (that includes direct effects of sex, SES, and ethnicity along with the rubric factors) fit the data better than a rubric model that is limited to the three rubric factors?

4. If so, what are the relative associations of those additional, non-rubric factors with grades?

Method

Participants

Teachers and students for this study were drawn from the participants in a large-scale experimental evaluation of Instructional Consultation Teams (Vu et al., 2011; Shanahan et al., 2011; Berger et al., 2011). The experimental study had four data collection waves: Pre-intervention baseline (2005-06), Year 1 intervention (2006-07), Year 2 intervention (2007-08), and Year 3 intervention (2008-2009). Data for the current study came from the 2008-09 school year. A study which examined effects of the IC Teams intervention found small and generally non-significant effects on student grades for the 2008-09 year (Shanahan et al., 2011).

Participants for the current study come from 45 schools in a suburban mid-Atlantic county and consist of students in grades 4 and 5 (N = 8,887), and their classroom teachers (N = 349). Table 1 details participant characteristics. Of the 45

schools, 34 participated in the experimental study of IC Teams. The remaining 11 schools, from the same school district, provided data as part of a separate quasi-experimental evaluation of the same intervention.

The participating county school system includes mostly suburban, but also some rural communities. In 2011, Newsweek magazine ranked all of the high schools in the system in the top 6% of high schools in the country (America's Best High Schools, 2011).

Procedures

Data for this study were collected through two methods. Student and teacher demographic information, student report card grades, and student standardized test scores were provided by the school system. Student behavioral information was collected using a Teacher Report on Student Behavior (TRSB) survey. The TRSB was administered via school system intranet in February of four consecutive years (2006-09) and was managed and monitored by school personnel. School officials requested that classroom teachers complete a TRSB report for each student in their classroom.

Measures

As explained in this section, the externalizing behavior scale from the TRSB survey provided teacher ratings of student problem behavior, and the concentration/learning scale provided teacher ratings of student academic effort. School system archival files provided student report card grades, standardized achievement test scores, and demographic information. Table 2 summarizes descriptive statistics and reliabilities for the measured variables, all of which are

shown in standardized score form ($M = 0$, $SD = 1$). Table 3 reports correlations between all measured variables in the study sample.

Student Variables

Student report card grades (outcome). Fourth and fifth grade students receive grades ranging from "F" (failure) to "A" (outstanding). Teachers in the participating school system assign grades according to a detailed rubric. This rubric outlines levels of performance (academic and effort) that correspond to each grade (see Appendices A and B). The rubric instructs teachers to consider the student's "achievement in subject," "class performance," and "independence in work" when assigning grades. The "achievement in subject" category includes mastery of academic material and class objectives. The "class performance" category includes participation, effort, neatness, timely submission of work, attendance, and originality of thinking and expression. The "independence in work" category includes self-direction, completion of independent work in addition to required assignments, timely completion of work, and need for encouragement to complete tasks. Although the rubric does instruct teachers to incorporate student "independent" and "performance" behaviors into their grading, it does not instruct them to incorporate conduct or behavior problems. Students receive a single omnibus grade for each subject area.

Report cards include grades for math, reading, science, social studies and writing across four marking quarters, and a final year-end grade for each subject area. Student grades were converted to numeric values using the following conversion: A=4, B+=3.4, B=3, C+=2.4, C=2, D+=1.4, D=1, F=0. Grades were averaged across the four marking periods for each subject to create subject specific GPAs. These

subject GPAs served as indicators of the report grades latent construct. Construct reliability for this variable was high (H = .93; Hancock & Mueller, 2001)

Monotonic transformations were attempted to address the negative skew of the report card grades distribution. Transformations attempted included square root, $\log_{10}$, ln, and inverse transformations. Scores were reflected and added to a constant (raising the lowest score to 1) prior to applying the transformations (Kline, 2010). None of these transformations appeared helpful. The ceiling effect for this variable could not be addressed through transformations, and led to a persistently skewed distribution.

Student academic achievement. The state achievement test in Virginia is known as Standards of Learning (SOL; Virginia Standards of Learning Assessments, 2009). Scores for the 2008-09 school year measured current student academic achievement. The SOL assesses achievement in reading, math, history (4^{th} grade only) and science (5^{th} grade only), and reports standard scores for each test subject. A composite achievement score was calculated for each student by averaging the standard scores earned on the SOL subject tests. For 4^{th} graders, this achievement composite was the average of reading, math, and social studies standard scores. For 5^{th} graders the composite was the average of reading, math, and science standard scores. In the current study, academic achievement is assumed to be a stable construct. Although students took the achievement tests in the spring, their performance is assumed to provide an estimate of the achievement levels they demonstrated throughout the year.

Several monotonic transformations were performed to normalize the achievement variable distribution. The square root transformation appeared to provide the best fix for the negatively skewed distribution. Scores were multiplied by negative one, and then added to a constant to increase all values to positive values. Next, the square root transformation was applied. Finally, scores were transformed back to their original direction so that high values represented high achievement.

Problem behavior. Externalizing behavior scale scores from the 2009 TRSB survey measured student problem behavior. The externalizing behavior scale consists of 8 items measuring the degree to which students are able to regulate their behavior, emotions, and interactions with other people. This scale was adapted for the present research from the TOCA-R (Werthamer-Larsson, Kellam & Wheeler, 1991). Modifications were minor and involved re-wording several items which seemed unclear, and reducing the number of response options from six to four. Sample items are "Takes others' property without permission," "Is physically aggressive or in fights with others," and "Defies teachers or other school personnel." Items were rated on a four point Likert-type scale (0=Never/Almost never, 1=Sometimes, 2=Often and 3=Very Often). Appendix C presents the full scale. Exploratory factor analysis (using PAF) of the adapted scale in the current sample generated a 1-factor solution, with loadings ranging from .563 to .828.

Scores on the problem behavior measure were recoded to a dichotomous form (0 = 0, all else = 1) to address the non-normality of the distribution of the original variable. Conceptually, this recoding procedure assigned '0' to students whose

teachers did not endorse any of the problem behavior items for them, and '1' to students who were rated as displaying some degree of problem behavior on the scale.

Student academic effort. Concentration/learning behavior scale scores from the 2009 TRSB survey measured student academic effort. The scale consists of eight items which measure the extent to which the child directs his or her undivided thought and attention toward the academic objective. Sample items include "Accomplishes assignments independently," "Eager to learn," and "Works to overcome obstacles in schoolwork." Items were rated on a 4 point Likert-type scale (i.e., 0 = Never/Almost never, 1 = Sometimes, 2 = Often and 3 = Very Often; adapted from the Teacher Observation of Child Adaptation, Revised, TOCA-R, measure by Werthamer-Larsson, Kellam, & Wheeler, 1991). Modifications to the scale, as with the problem behavior scale, were minor. Appendix C presents the full list of items for the effort scale used in the current study. Exploratory factor analysis (using PAF) of the adapted scale in the current sample generated a 1-factor solution, with loadings ranging from .505 to .877.

Scores on the concentration scale were recoded (>2 = 1, <=2 = 0). Conceptually, this means that students who were rated as displaying low levels of concentration were coded '0' while students who received high concentration ratings were coded '1'.

Attendance. The attendance variables is the number of days a student was absent in the 2008-09 school year, recoded to a dichotomous variable to address non-normality of the distribution and outliers. Students with five or fewer absences were coded '0.' Students with greater than five absences were coded '1.'

Student demographic characteristics. Student sex (0=Female, 1=Male), ethnicity (White = 1, All else = 0), and Free and/or Reduced Meals (FARM) status (1=receiving FARM, 0=not receiving FARM) were categorical variables describing student demographic characteristics.

Student grade level. To address concerns that grading practices could differ across grade level, student grade level was included as a predictor in both the rubric and non-rubric models (1=grade 5, 0=grade 4). This variable was included to control for variance associated with grade level.

Technical Concerns

Centering of variables. Dichotomous variables (effort, attendance, problem behavior, sex, white and FARM) remained uncentered. Student achievement was grand-mean centered. These centering decisions have implications for interpretation of the intercept (Heck, 2009). The intercept is the outcome for a student whose value for the achievement variable equals the grand mean for the sample, and whose value for dichotomous variables equals zero.

Multicollinearity. Multicollinearity is pervasive across many statistical methods including multilevel modeling and structural equation modeling (Heck, 2009; Kline, 2011; Grewal, Cote & Baumgartner, 2004). Multicollinearity occurs when there are high correlations among variables (often among predictors). When multicollinearity occurs to a high degree, it can have adverse effects such as causing the analysis to fail or decreasing the precision of regression estimates (Heck, 2009; Kline, 2011; Grewal, Cote & Baumgartner, 2004). Ideally, multicollinearity

problems should be addressed prior to data analysis, before problems are even observed.

Several steps were taken to reduce the risks inherent in models exhibiting a high degree of multicollinearity. Correlations were examined prior to including variables in the model to determine the presence of extremely high intercorrelations (greater than .95) or moderately high correlations (greater than .6; Grewal et al. 2004). None of the correlations fell within the extreme or high range, so all of the variables were included in subsequent analyses. In addition to examining correlations between variables, reliabilities for the factor indicators were also examined. Grewal et al. recommended that researchers insure that factor indicators have high reliability because high reliability leads to more accurate estimation and fewer Type II errors. Reliabilities (coefficient alpha) of the indicators for the report card grades construct were all high (greater than or equal to .90).

During the analysis of the data, coefficients and their standard errors were monitored as additional variables were added to subsequent models. Originally, an English Speakers of Other Languages (ESOL) variable was included in the analysis, but it was removed for the final analysis because its coefficient and standard error changed undesirably when it was added to the models.

Missing data. Researchers across academic disciplines routinely encounter missing data, and seek ways to address missing data problems in their analyses (Enders, 2010). Different analytic approaches are recommended for dealing with missing data depending on the cause of the missing data (Enders, 2010). For this reason, missing data were identified and examined prior to data analyses.

Examination of the data in the current study showed that the student report card grade, achievement score, sex, FARM, and Ethnicity variables had no missing data. Missing data was a problem, however, for the problem behavior and effort variables. About 11% of cases were missing ratings for problem behavior, and 11% were missing ratings for effort. In general, students who were missing a problem behavior rating were also missing a rating for effort. Missing data were expected for the problem behavior and effort variables as these data were collected through a teacher survey. Although 94% of teachers completed the survey in the outcome year, some teachers did not complete reports for all of their students despite being instructed to do so. Missingness did not appear related to the outcome variable (report card grades) or to the achievement or demographic variables.

In addition to the missing data patterns described above, it was also important to consider how the probability of missing data might be related to the measured variables (Enders, 2010). The following three missing data mechanisms were considered: missing not at random (MNAR), missing at random (MAR), and missing completely at random (MCAR). These mechanisms are based on Rubin's (1976) classification of missing data problems, and were recently described by Enders (2010).

The observed data did not seem to represent a random sample of observations (because students tended to be missing scores for problem behavior *and* effort), and therefore did not seem to meet the criteria for the missing completely at random (MCAR) mechanism. It was possible that data were missing not at random (MNAR), but this seemed unlikely because it would mean that teachers completed

behavior ratings, or failed to complete them, based on their perceptions of their students' behaviors. It seemed more likely that behavior scores were missing due to teacher characteristics such as their attitudes towards the research study, attitudes towards completing lots of questionnaires, or simply their perception of the time that would be required to complete the reports.

The missing at random (MAR) data mechanism seemed to provide a reasonable explanation for the missing data concerns in the current study. For data to be MAR, the probability of missing data on a variable can be related to other measured variables, but cannot be related values on the variable itself. Unfortunately, there is no formal test to determine whether or not the MAR assumption is operating. In addition, there are no formal tests to rule out the possibility that data are MNAR. The final decision to move forward treating the missingness as MAR was informed by examining missing data patterns across the study variables.

Enders (2010) recommended using maximum likelihood estimation to work with data that are MAR because maximum likelihood "yields unbiased parameter estimates with MAR data" (Enders, 2010, p. 87). Although there is no perfect way of coping with missing data, maximum likelihood methods generally reduce potential bias in model parameters. Analysis procedures operated on the assumption that data were MAR.

Outliers. Examination of potential outliers occurred prior to any formal modeling, and indicated outliers on several variables. Outlier detection is an important aspect of any data analysis that assumes variables follow a normal distribution. If unattended, these outliers create non-normality in data distributions,

which can threatened the validity of inferences drawn from statistical analyses which rely on this distributional assumption. Outliers for the problem behavior, effort, and attendance variables were addressed when those variables were recoded into categorical variables. Outliers on the report card grade and achievement variables were identified by calculating modified z-scores and removing cases with modified z-scores greater than 3.5. In total, about 65 cases were deleted (less than .01% of the original student population).

Data Analysis

Analyses addressed both the multilevel nature of the data (students clustered within classrooms) and the hypothesized structural relations using a multilevel structural equation modeling approach. All structural equation models were estimated using the MLR estimator and type=complex analysis command within the Mplus software program. The MLR estimator is the only estimator available for complex samples (e.g. nested data) when dependent variables are all continuous (Muthén, 2010).

In general, all maximum likelihood estimation methods work by trying out different population parameter values, and finally settling on the estimates with the greatest likelihood of having produced the actual sample data (Enders, 2010). In this way, maximum likelihood searches for the population parameters with the maximum likelihood of producing the sample. MLR (which was used in the current study) is a type of maximum likelihood estimation that computes robust standard errors instead of the conventional model-based standard errors computed by tradition ML.

The analysis approach used in the current study used a sandwich estimator to compute robust standard errors for the variance components of the estimates at both the student and classroom levels (Muthén, 2010). Sandwich estimators are used to correct the standard errors when evidence suggests that residuals are not normally distributed (Maas & Hox, 2004). Without this correction, the variance components for the regression coefficients can be biased. The type=complex analysis option in Mplus automatically computed these robust standard errors and did not require the researcher to specifically request this output.

Data analyses proceeded in stages. First, the measurement model for the latent construct (report card grades) was defined. Rather than estimating the common factor model at each data level, the factor structure was assumed to be the same at both levels. Conceptually, this approach implies that the report card grades construct is the same across both the student and classroom levels, and leads to development of a variance components measurement model which simplifies the measurement model, simplifies interpretation of the latent construct, and is recommended whenever possible (N. C. Gottfredson, 2008). According to N. C. Gottfredson (2008, p. 44) "a MSEM [multilevel structural equation model] would be extremely difficult to implement if the between- and within-measurement model formulation were used." Figure 1 shows the proposed variance components measurement model for the report card grades latent construct.

Next, the structural models were specified. In these models, the latent construct (report card grades) was regressed on student-level measured variables (achievement, effort, attendance, problem behavior, male, FARM, and White). The

hypothesized model (non-rubric model) is shown in Figure 2. An alternative model (rubric model), nested within the hypothesized model, is shown in Figure 3, with paths from non-rubric factors fixed to zero. Paths are marked with a + or – sign to indicate the direction of the hypothesized effect. Variables that vary across both students and classrooms are shown inside the box. Error influences at the classroom level are shown outside the box, as they vary only at the classroom level.

Then, the measurement model and structural models were estimated simultaneously at the student-level, ignoring clustering, to see if specification errors occurred at that level (Kline, 2011). Finally, both the student-level and classroom-level models were simultaneously estimated for both the non-rubric model (hypothesized) and rubric model (alternative).

Difference testing compared the rubric model to the non-rubric model to determine which model provided the superior fit for the data. This difference test was conducted using loglikelihood values and scaling correction factors generated by the MLR estimator (Asparouhov & Muthén, 2010; Satorra & Bentler, 2001; Chi-Square Difference Testing Using the Satorra-Bentler Scaled Chi-Square). Loglikelihood values and their corresponding scaling correction factors were used (rather than chi-square values) because chi-square values were not available for the models. According to Muthén (September 28, 2009), "chi-square and related fit statistics are not available when means, variances, and covariances are not sufficient statistics for model estimation. Nested models can be tested using -2 times the loglikelihood difference which is distributed as a chi-square." Difference testing followed this recommendation by Linda Muthén, one of the Mplus program developers.

Results

ICC for Grades

The intraclass correlation (ICC) for report card grades was .18. This ICC value means that 18% of the variance in the grades variable was between classrooms, and suggested the variance between classrooms was high enough to warrant multilevel approaches for modeling the report card grades construct (Heck, 2009). For latent variables, the ICC is the estimated variance at the between level divided by the total variance for the latent variable (Muthén, 1991).

Measurement Model

Data-model fit indices indicate that the measurement model is a reasonable representation of the relations that underlie the data ($\chi^2 = 7.964$ ($df = 2$, $p = .019$), RMSEA = .019, CI_{90}: (.006, .033) and CFI = .999).

Parameter Estimates

Parameter estimates for the non-rubric model are shown in Table 4 and Figure 4. In general, all of the variables were significantly associated with grades, and the associations were in the hypothesized direction. Effect estimates were medium-sized for achievement (.57) and effort (.30), but small (less than .10) for all other variables (Cohen, 1988). Individually, the effect estimates were all small for the non-rubric factors. The *R*-squared for the grades latent construct was .66.

Parameter estimates for the rubric model are shown in Table 5 and Figure 5. The pattern of associations for achievement, effort, and attendance mirrored those identified in the non-rubric model, and effect estimates were comparable, but slightly higher. The *R*-squared value for the grades construct was .64.

Information criteria were somewhat smaller for the non-rubric model (AIC = 207,218.905, BIC = 207,670.4) than the rubric model (AIC = 207,533.3, BIC = 207,956.6). Smaller AIC and BIC values imply that the non-rubric model provides a better fit for the data. Both the AIC and BIC fit indices point to the non-rubric model as the preferred model.

According to Kuha (2004), "model choice is easiest when AIC and BIC agree on the preferred model. This is then unlikely to be far from the best of the candidate models. Such a choice is also very robust in the theoretical context of both AIC and BIC" (Kuha, 2004; pp. 223-224). For this reason, both AIC and BIC values are reported here, even though they have different theoretical underpinnings and different aims. When using the BIC criterion, the aim is to identify the model that is most likely the true model for the data. In contrast, the aim when using the AIC is to identify the model that will do the best job at predicting future data (Kuha, 2004).

Difference Testing

Results from the difference test supported rejection of the null hypothesis that the two models fit the data equally well, and suggested that the non-rubric model provided a superior fit (TRd = 174.5; $df = 4$, $p < .001$).

Discussion

Results of this study examine factors that are associated with student report card grades to explore whether teachers in a large suburban school district follow the district grading rubric when assigning grades. If teachers are following grading rules, then a model which includes only rubric-specified factors should fit the data better than a model that includes non-rubric factors.

Report card grades appear to be complex, accounted for by many factors. But the most important factors, according to the results of the current study, seem to be academic achievement and effort, which are the two factors that teachers are instructed to incorporate into their grading. This finding is encouraging, because it suggests that although grades are not a perfect, clean measure of the rubric factors, they are mostly accounted for by those rubric factors.

In addition, results are consistent with prior research on grading practices, which consistently finds that teachers consider academic achievement more than any other factor when assigning grades (Brookhart, 1994; Bursuck, Polloway, Plante, Epstein, Jayanthi, & McConeghy, 1996; Cross & Frary, 1996; McMillan, Myran & Workman, 2002; Strein & Meshbesher, 2006). The current study found that the achievement variable had by far the largest association with grades.

When a rubric-only model of grades was compared to a model that included non-rubric factors, the more complex model seemed to provide a better fit for the data. The model that included the additional factors, which are theoretically related to grades, did a better job accounting for the variance in grades than a model with just the rubric factors. These results suggest that report card grades *are* due in part to variables outside the grading rubric.

Model comparison results, however, must be interpreted in terms of their practical significance for explaining grading practices. Effect estimates for the additional factors (such as problem behavior, ethnicity, sex, and FARM) were small compared to effect estimates for the primary rubric factors (achievement and effort). Although the relation of ethnicity with grades appeared moderate when it was the

only predictor of grades (during preliminary analyses), this relation declined after controlling for achievement and effort in the final analyses. The relation of ethnicity with grades, therefore, seems mostly accounted for by achievement and effort, and does not seem to be due to teacher biases.

Small effect sizes for separate demographic characteristics could underestimate the true association of these constructs with grades. Students who fit just one of the demographic categories might tend to experience just a small change in grades that is not practically significant. Students who are exposed to multiple risk factors might experience greater educational disadvantage, corresponding to larger negative associations with grades (Novotny, 2011; Boado, 2011). Supplemental analysis computed an "academic disadvantage" variable that included the following demographic characteristics: Ethnicity (0=White or Asian, 1=All other ethnicities), FARM (0=not receiving FARM, 1=receiving FARM), and ESOL (0= not receiving English Language Learning Services, 1= receiving English Language Learning services,). Student scores on the three demographic variables were summed to create the disadvantage variable, with values ranging from zero to three. For example, a Hispanic student who received FARM and ESOL services would be coded as "3" on the variable, while an Asian student not receiving FARM or ESOL would be coded "0." A supplemental non-rubric model of grades included this disadvantage variable in place of the white and FARM variables. Effect estimates for achievement (.54) and effort (.31) remained consistent with the main analysis, while the effect estimate for disadvantage (-.12) was slightly stronger than the effects for the individual demographic characteristics on their own. Although the magnitude of the association

with grades increased slightly for the disadvantage variable, the increase was not as much as might be expected given the research on disadvantage and educational attainment.

Disadvantage may have larger associations with grades when student-level data are aggregated to the classroom level. The main study used disaggregated, student-level data, and found small associations of demographic characteristics with grades, but these associations might be much larger at the classroom level. Slight effects at individual levels can produce strong associations at the school and class levels as these slight biases cumulate to produce large group effects. Researchers across scientific disciplines consistently find this aggregation bias (e.g. Finney, Humphreys, Kivlahan & Harris, 2011; Monteforte, 2006; Ouimet, 2000).

Supplementary analyses examined possible effects of aggregation on relations among grades and other variables in the current sample. Student data were aggregated to the classroom level and entered into multiple regression equations (see Appendix E). Effect estimates for classroom achievement, classroom effort, and classroom disadvantage were nearly equal in size (.25 to .29). These results suggest a lower association of average classroom achievement with average classroom grades, and a much higher association of disadvantage with grades when data are aggregated to the classroom level. A more sophisticated analysis might employ multilevel modeling, or multilevel SEM, to examine associations at the classroom level to see if these classroom level associations are diminished once the characteristics and behavior of individual students are taken into account. However that may be, this examination indicates that although the contributions of non-rubric factors to individual students'

grades are small, these small influences may produce larger associations at the classroom level.

Although results suggest that overall, teachers appear to follow the grading rubric when assigning report card grades, it is possible that some teachers could veer from this general trend. Classroom or teacher characteristics could moderate the within-classroom slopes of the predictors with grades. This possibility was explored using two-level hierarchical linear models, with students nested within classrooms, to test heterogeneity of slopes of grades with rubric and non-rubric variables. Results indicated significantly varying slopes for achievement, effort, problem behavior, male, FARM, and disadvantage. Slopes did not significantly vary for attendance or White. Future research could explore the sources of differences among teachers in what influences their grades. , The implications of these teacher differences for student outcomes might prove useful, but is beyond the scope of the present study.

Potential Limitations

Multilevel SEM provides an exciting analysis tool, but remains a new and still developing approach that is not well understood (Kline, 2011; Heck, 2009). Multiple regression analysis served as a sensitivity test for the main analysis approach. Predictor variables were added one by one using a hierarchical regression procedure (starting with rubric variables). Standardized coefficients for all variables were consistent with the coefficients generated by the multilevel SEM analysis employed in the main study (See Appendix D), generating support for the main findings.

Ambiguous temporal precedence of the measured variables threatens the ability to make inferences about the direction of their associations with report card

grades. The current study proposed that measured variables such as achievement and effort influence grades, but it is possible that the report card grades construct is not a completely endogenous variable, and may affect other variables in my structural model (Kline, 2011). In fact, one could argue that grades and student behaviors may have reciprocal effects (feedback loop). For example, Bonesrønning (2004) found that hard grading practices led to increased student achievement. Here, I limit my focus to direct effects of rubric and non-rubric factors on grades, but future studies could examine reciprocal effects of such variables.

Specification errors, such as leaving out unanalyzed associations, omitting correlated residuals suggested by theory, or omitting causal variables correlated with variables in my model, could also threaten causal inference. However, the standard that must be met in the current study is lower that would be required for confident causal inference, given the scope of the research questions. The current study is descriptive in purpose, and does not intend to examine a complete causal model of grades.

As with most social science research, it is possible that the variables I use in this study do not measure the constructs that I intend to measure. Student problem behavior and academic effort measures are teacher reports. These reports represent teacher perceptions of student behavior, and do not necessarily capture only the actual incidence of problem behavior or effort. Shared method variance for report card grades, effort, and problem behavior, which are all teacher reports, could results in halo effects for these variables.

Construct validity for the demographic variables is also questionable. These variables only indicate student membership in one of the categories, and the categories are not intended in the present research to have any well-defined meaning. For example, the FARM variable indicates whether students receive free and/or reduced meals, but cannot capture the complex factors that contribute to the construct of SES. Similarly, the ethnicity variable was coded 1=White, 0=all other ethnicities to simplify the models, but effects for this variable cannot be meaningfully interpreted beyond the fact that the variable explains variance in grades.

The results are based on a sample of fourth and fifth grade students and teachers in a suburban public school system in the mid-Atlantic region and are specific to the demographic characteristics unique to this school system, such as SES percentages, ethnicity classifications, or teacher education. Results are of unknown generalizability to students younger or older than those in this study, or to teachers of other grade levels. This study only includes students from grades 4 and 5 because in this school system, the standardized achievement tests are only give to elementary students in third, fourth, and fifth grade. Future studies might examine K-2 and 6-12 grade students and teachers.

Finally, skewed distributions for the report card grade variables violate the assumption of normality, which threatens statistical conclusion validity. Transformations to normalize distributions were attempted, but did not appear helpful for the grades variables due to strong ceiling effects.

Implications and Future Directions

Results of the current study have implications for understanding how teachers assign grades, and how well student grades reflect the categories outlined on grading rubrics. In addition, results address a gap in our current understanding of grading practices and grading policies.

Grades are often overlooked as achievement measures because they are "contaminated" with additional information, such as behavior, effort, or biases (Allen, 2005; Greville, 2009). Results from this study clarify the factors that contribute to this contamination, and have implications for educational researchers searching for valid measures of academic achievement. Overall, results suggest that concerns with grading "contamination" maybe be exaggerated. Most of the variance in grades in the current sample is accounted for by the factors outlined on the district's grading rubric.

It seems prudent to investigate the factors that contribute to grades any time that grades are used in educational research. In addition, researchers would be wise to examine the grading rubrics that correspond to the report card data they are using in research studies.

Future research might ask teachers in the participating school district to describe their grading practices, or to respond to a self-report questionnaire, similar to McMillan et al. (2002), to indicate the factors they consider when grading. Results from such an interview or questionnaire study could compliment the current study by showing how teachers perceive their own practices, and whether that perception is consistent with measured relations of grades with rubric and non-rubric variables.

Results provide a clearer picture of the factors that contribute to grades for 4th and 5th graders in the participating school district, but might differ in school systems employing different grading rubrics. The current study could be replicated in other samples to determine how well grades in other districts reflect the stated grading policies. Findings from the current study may be specific to the participating district, or might reflect an overall pattern that applies to many districts. Other research already shows that grading practices vary widely across districts, but we do not know if this variability is related to differences in grading rubrics, differences in teacher grading practices, or other factors. Research investigating grading practices and the policies that guide them is just beginning.

Table 1
Participant Characteristics

Teachers (N = 374)	Percentage	Students (N = 8,555)	Percentage
Gender		Gender	
Female	85	Female	49
Male	15	Male	51
Ethnicity		Ethnicity	
Caucasian	83	Caucasian	43
African-American	10	African-American	21
Hispanic	3	Hispanic	24
Asian	1	Asian	7
Other	3	Other	3
Grade Level		Grade Level	
4th grade	51	4th grade	51
5th grade	49	5th grade	49
Age		FARM	34
23-30 years	30	ESOL	22
31-40 years	30	Special Education	12
41-50 years	17		
51-60 years	18		
61 and older	5		

Note. FARM is Free and/or reduced meals. ESOL is English speakers of other languages.

Table 2
Descriptive Statistics

	Raw Scores				Standardized (z) Scores				
Variables	Mean/ proportion	*SD*	Min	Max	Mean/ proportion	*SD*	Min	Max	Reliability
Math GPA	3.12	.71	.50	4.00	.00	1.00	-3.67	1.26	.91
Reading GPA	3.12	.67	.67	4.00	.00	1.00	-3.63	1.33	.90
Science GPA	3.17	.68	.75	4.00	.00	1.00	-3.57	1.22	.90
Social Studies GPA	3.18	.70	.67	4.00	.00	1.00	-3.61	1.17	.90
Achievement	492	62	266	600	.00	1.00	-2.60	2.77	.82
Effort	.56	.49	.00	1.00	.00	-	-1.13	.88	.92
Attendance	.48	.50	.00	1.00	.00	-	-0.97	1.04	-
Problem Behavior	.50	.50	.00	1.00	.00	-	-1.00	1.00	.90
Male	.50	.50	.00	1.00	.00	-	-1.00	1.00	-
White	.43	.49	.00	1.00	.00	-	-.87	1.15	-
FARM	.33	.47	.00	1.00	.00	-	-.70	1.43	-

Note. Reliability is coefficient α (alpha) measuring internal consistency. Descriptive statistics include grades 4 and 5. Math, reading, science, and social studies GPAs are the average of grades across 4 marking periods in that subject area. Problem behavior is the score a student received for the externalizing behavior scale on the TRSB survey, recoded to a dichotomous variable. Effort is the score a student received for the concentration/learning scale on the TRSB survey, recoded to a dichotomous variable. Reliabilities for problem behavior and effort are reported for the original scale, prior to recoding. FARM is receiving free and/or reduced meals. Achievement is the average of student achievement scores across the math, reading, science and social studies tests. Attendance is the number of absences a student had in the school year, recoded to a categorical variable.

Table 3
Correlations Between Variables

Variables	1	2	3	4	5	6	7
1 Achievement	-						
2 Effort	0.42***						
3 Attendance	-0.04***	-0.06***					
4 Grade 5	-0.02	-0.03	-0.02*				
5 Problem Behavior	-0.19***	-0.39***	0.01	0.04***			
6 Male	-0.03*	-0.18***	0.02*	-0.01	0.13***		
7 White	0.29***	.11***	0.11***	0.01	-0.05***	0.01	
8 FARM	-0.34***	-.14***	0.01	-0.01	0.06***	0.01	-0.43***

Note. *p < .05. **p <.01. ***p < .001.

Table 4

Parameter Estimates for Hypothesized and Significant Paths from the "Non-Rubric" Model

Outcome	Predictor	Estimate (SE)	Standardized Effect	*p*-value
Grades				
	Achievement	.105 (.003)***	.550	<.001
	Effort	.382 (.017)***	.310	<.001
	Attendance	-.057 (.009)***	-.047	<.001
	Grade 5	-.026 (.025)*	-.021	.300
	Problem Behavior	-.045 (.013)***	-.034	.001
	Male	-.085 (.010)***	-.070	<.001
	White	.069 (.013)***	.056	<.001
	FARM	-.111 (.015)***	-.085	<.001
Achievement with Effort		.666 (.021)***	.425	<.001
Achievement with Attendance		-.070 (.017)***	-.044	<.001
Achievement with Grade 5		-.084 (.037)*	-.017	.021
Achievement with Problem Behavior		-.313 (.021)***	-.191	<.001
Achievement with Male		-.033 (.018)*	-.025	.064
Achievement with White		.451 (.021)***	.289	<.001
Achievement with FARM		-.502(.020)***	-.344	<.001
Effort with Attendance		-.014 (.003)***	-.057	<.001
Effort with Grade 5		-.007 (.005)	-.029	.181
Effort with Problem Behavior		-.096 (.004)***	-.388	<.001
Effort with Male		-.045 (.003)***	-.180	<.001
Effort with White		.028 (.003)***	.114	<.001
Effort with FARM		-.032 (.003)***	-.136	<.001
Attendance with Grade 5		-.006 (.003)*	-.024	.047
Attendance with Problem Behavior		.002 (.003)	.008	.496
Attendance with Male		.005 (.003)	.022	.067
Attendance with White		.027 (.003)***	.107	<.001
Attendance with Farm		.002 (.003)	.010	.391
Grade 5 with Problem Behavior		.010 (.005)	.040	.060
Grade 5 with Male		-.002 (.002)	-.007	.353
Grade 5 with White		.002 (.007)	.007	.801
Grade 5 with FARM		-.003 (.007)	-.011	.684
Problem Behavior with Male		.033 (.003)***	.133	<.001
Problem Behavior with White		-.014 (.004)***	-.056	<.001
Problem Behavior with FARM		.013 (.003)***	.057	<.001
Male with White		.002 (.003)	.010	.338
Male with FARM		.002 (.002)	.010	.291
White with FARM		-.100 (.003)***	-.430	<.001

Note. *$p < .05$. **$p < .01$. ***$p < .001$. AIC = 207,218.9. BIC = 207,670.4. Sample-size adjusted BIC = 207,467.0.

Table 5
Parameter Estimates for Hypothesized and Significant Paths from the "Rubric" Model

Outcome	Predictor	Estimate (SE)	Standardized Effect	*p*-value
Grades				
	Achievement	.112 (.003)***	.591	<.001
	Effort	.417 (.016)***	.338	<.001
	Attendance	-.049 (.009)***	-.040	<.001
	Grade 5	-.022 (.025)	-.018	.395
Achievement with Effort		.666 (.021)***	.416	<.001
Achievement with Attendance		-.070 (.017)***	-.044	<.001
Achievement with Grade 5		-.084 (.037)	-.052	.021
Achievement with Problem Behavior		-.313 (.021)***	-.194	<.001
Achievement with Male		-.033 (.018)*	-.021	.064
Achievement with White		.451 (.021)***	.283	<.001
Achievement with FARM		-.502 (.020)***	-.331	<.001
Effort with Attendance		-.014 (.003)***	-.057	<.001
Effort with Grade 5		-.007 (.005)	-.029	.191
Effort with Problem Behavior		-.097 (.004)***	-.389	<.001
Effort with Male		-.046 (.003)***	-.184	<.001
Effort with White		.029 (.003)***	.118	<.001
Effort with FARM		-.033 (.003)***	-.141	<.001
Attendance with Grade 5		-.006 (.003)*	-.024	0.047
Attendance with Problem Behavior		.002 (.003)	.008	0.485
Attendance with Male		.005 (.003)	.022	0.067
Attendance with White		.027 (.003)***	.107	<.001
Attendance with Farm		.002 (.003)	.010	0.391
Grade 5 with Problem Behavior		.010 (.005)	.041	0.058
Grade 5 with Male		-.002 (.002)	-.007	0.353
Grade 5 with White		.002 (.007)	.007	0.801
Grade 5 with FARM		-.003 (.007)	-.011	0.684
Problem Behavior with Male		.034 (.003)***	.134	<.001
Problem Behavior with White		-.014 (.004)***	-.057	<.001
Problem Behavior with FARM		.014 (.004)***	.059	<.001
Male with White		.002 (.003)	.010	0.338
Male with FARM		.002 (.002)	.010	0.291
White with FARM		-.100 (.003)***	-.430	<.001

Note. *p < .05. **p <.01. ***p < .001. AIC = 207,533.3. BIC = 207,956.6. Sample-size adjusted BIC = 207,765.9

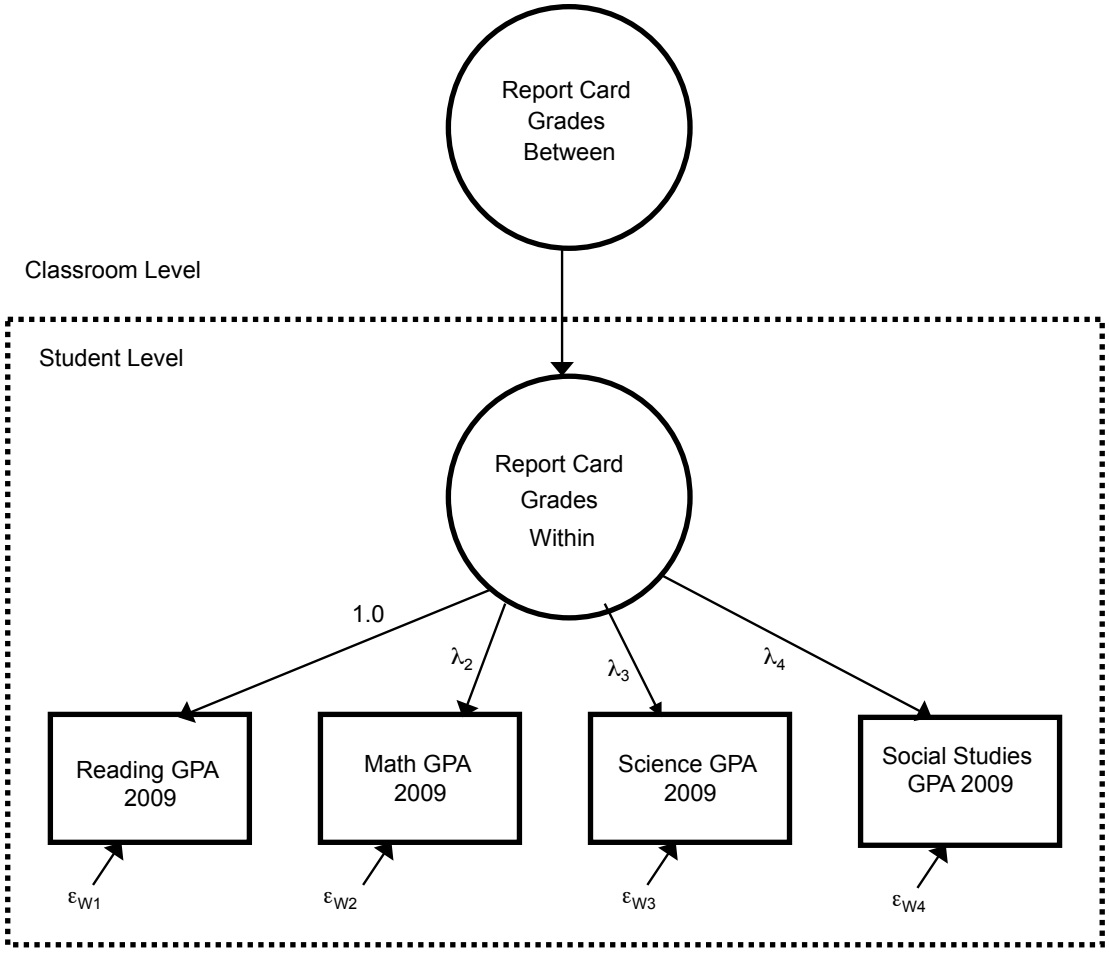

Figure 1. Hypothesized measurement model for report card grades.

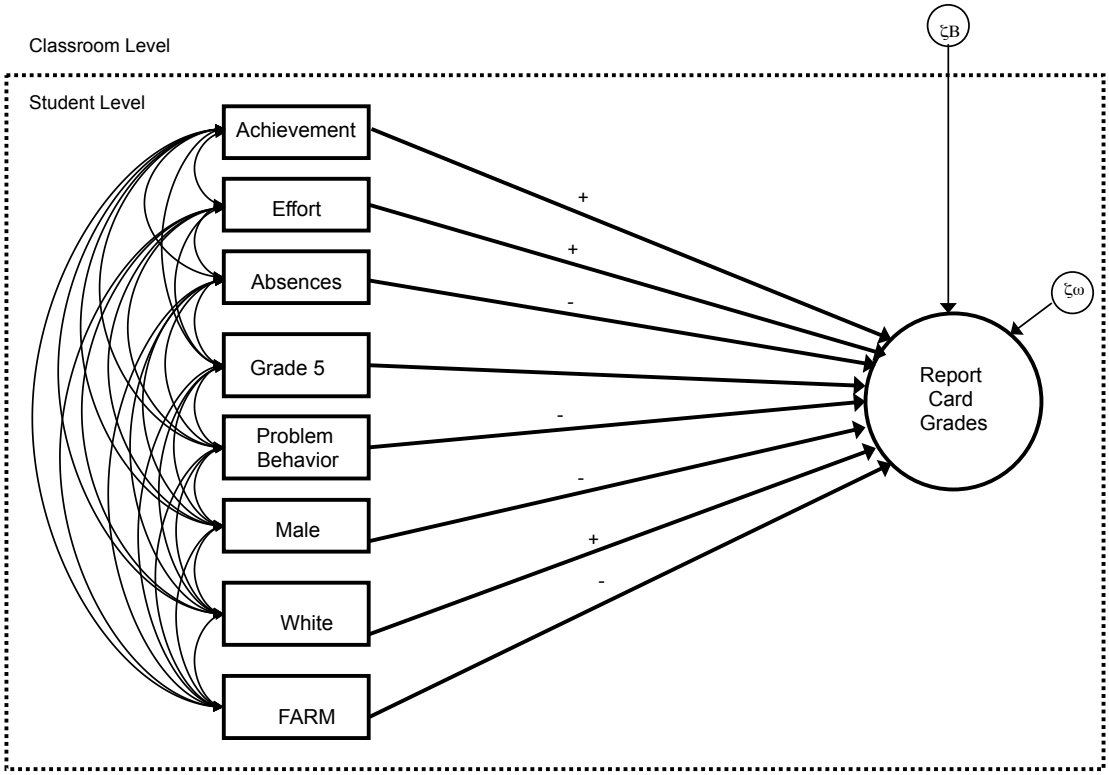

Figure 2. Hypothesized non-rubric model. Paths are marked with a + or – sign to indicate the direction of the hypothesized effect. Variables that vary across both students and classrooms are shown inside the box. For simplicity of presentation, error influences for measured variables are not shown (but would be present at both the student and classroom level).

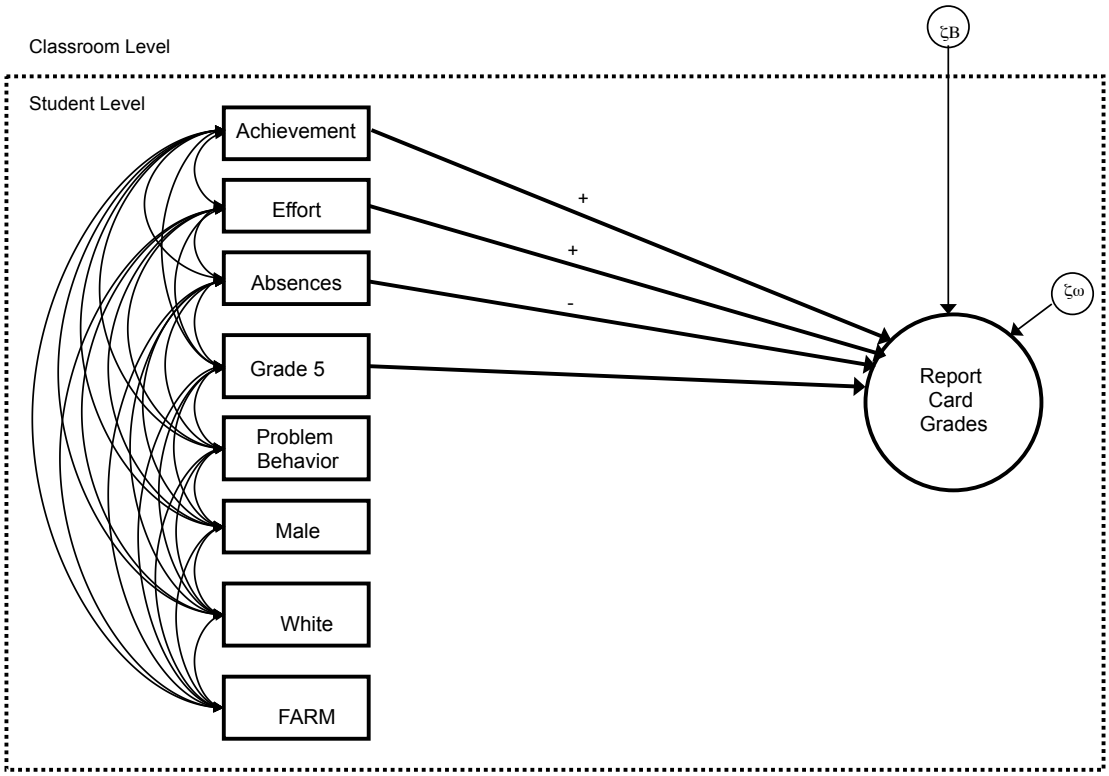

Figure 3. Hypothesized rubric model. Paths are marked with a + or – sign to indicate the direction of the hypothesized effect. Variables that vary across both students and classrooms are shown inside the box. For simplicity of presentation, error influences for measured variables are not shown (but would be present at both the student and classroom level).

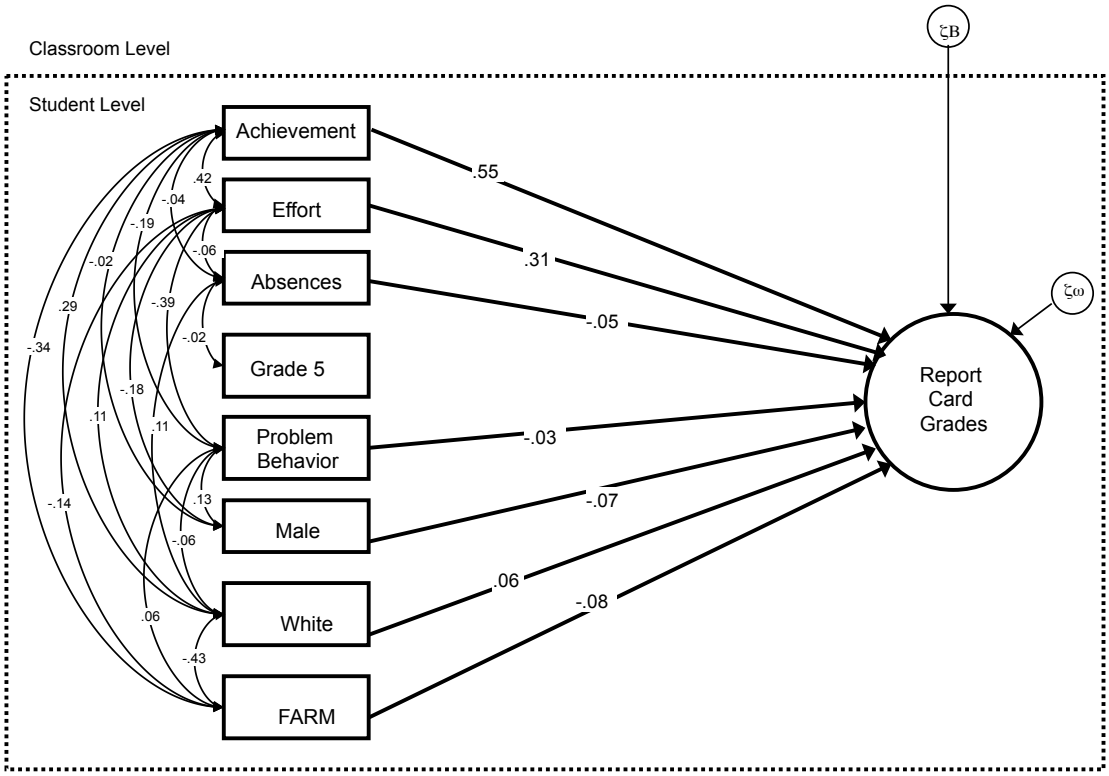

Figure 4. Final Non-rubric Model. Structural paths and correlations shown are statistically significant at $p < .05$ or lower. Structural paths are standardized. All exogenous variables are allowed to correlate. Non-significant correlations between exogenous variables are omitted. Variables shown inside the box vary at both the student and classroom level.

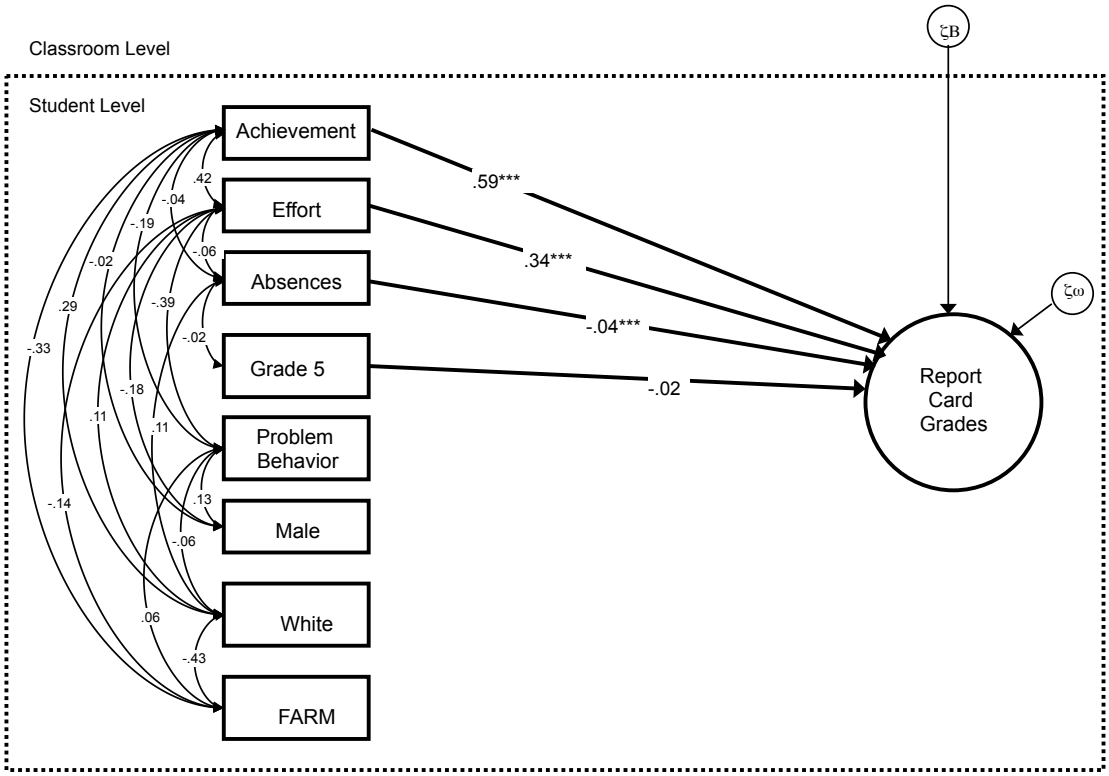

Figure 5. Final Rubric Model. *$p < .05$. **$p < .01$. ***$p < .001$. Paths shown are standardized. All exogenous variables are allowed to correlate. Structural paths not shown were fixed to zero. Non-significant correlations between exogenous variables are omitted. Variables shown inside the box vary at both the student and classroom level.

Appendix A

Grading Rubric for Grades 3-5 Core Subjects: A Through C+

Regulation 661-1, INSTRUCTION, June 23, 2004, Pages 3-5

GRADE	RANGE	ACHIEVEMENT IN SUBJECT	CLASS PERFORMANCE	INDEPENDENCE IN WORK
A Excellent	93 - 100 %	- demonstrates outstanding achievement and mastery of the subject area - evidences understanding and appreciation of the fundamental concepts of the subject area - exercises superior ability in problem solving and in arriving at logical conclusions - expresses ideas clearly both orally and in writing	- fully participates and demonstrates effort in all class activities - exhibits originality in thinking, expression, and work products - submits all work on or before due date - displays neatness, legibility, and accuracy in work	- is self-directed - shows originality in preparation of assignments - consistently contributes independent work in addition to required assignments - submits all work on or before due date
B+ Very Good	90 – 92 %	- demonstrates very good achievement and mastery of the subject area - evidences understanding and appreciation of the fundamental concepts of the subject area - expresses ideas clearly both orally and in writing	- usually participates and demonstrates effort in class activities - exhibits originality in thinking, expression, and work products - submits all work on or before due date - displays neatness, legibility, and accuracy in work	- completes assignments on time, thoroughly and accurately - is self-directed - sometimes contributes independent work in addition to required assignments
B Good	84 - 89 %	- demonstrates above average achievement and mastery - usually evidences understanding and appreciation of the fundamental concepts of the subject area	- usually participates and demonstrates effort in class activities - usually submits work on or before due date - displays neatness, legibility, and accuracy in work	- usually completes assignments on time, thoroughly and accurately - is self-directed - sometimes contributes independent work in addition to required assignments
C+ High Average	81 – 83 %	- achieves sufficient subject mastery to proceed to the next level - objectives are usually mastered, but not always	- sometimes participates and demonstrates effort in class activities - inconsistently submits work on due date - does not always display neatness, legibility, and accuracy in work	- usually completes assignments on time - is sometimes self-directed, but sometimes needs encouragement to complete tasks

Appendix B

Grading Rubric for Grades 3-5 Core Subjects: C through F

Regulation 661-1, INSTRUCTION, June 23, 2004, Pages 3-5

GRADE	RANGE	ACHIEVEMENT IN SUBJECT	CLASS PERFORMANCE	INDEPENDENCE IN WORK
C Average	74 – 80 %	- achieves sufficient subject mastery to proceed to the next level - objectives are sometimes mastered, but not always	- sometimes participates and demonstrates effort in class activities - inconsistently submits work on due date - does not always display neatness, legibility, and accuracy in work	- sometimes completes assignments on time - is sometimes self-directed, but sometimes needs encouragement to complete tasks
D+ Below Average	71 – 73 %	- frequently falls below the average level of achievement - lacks sufficient subject mastery to proceed to the next level	- often does not participate and demonstrate effort in class activities - submits poor work, but effort is in evidence	- frequently requires individual direction - often does not complete assignments on time, or at all
D Poor	65 – 70 %	- demonstrates limited achievement of grade level objectives - consistently falls below grade level requirements	- may be irregular in attendance and generally fails to make up missed work - shows little interest in class and rarely contributes	- seldom completes an undertaking without teacher direction and encouragement
F Failure	64 % and below	- fails to meet minimum requirements	- frequently fails to complete assignments - demonstrates little or no effort - may have excessive unexcused absences - fails to complete 65% of the assigned, evaluated work	- seldom completes an undertaking without teacher direction and encouragement

Appendix C

Teacher Report on Student Behavior

Variable	*Item*
	Concentration/Learning Scale
I_2R	Easily distracted (reverse score)
I_5	Accomplishes assignments independently
I_12	Eager to learn
I_15	Works to overcome obstacles in schoolwork
I_18R	Says things like "I can't do it" when work is difficult (reverse score)
I_21	Stays on task
I_23	Pays attention
I_24	Learns up to ability
	Externalizing Behavior Scale
I_4	Defies teachers or other school personnel
I_7	Argues or quarrels with others
I_9	Teases or taunts others
I_11	Takes others property without permission
I_13	Is physically aggressive or fights with others
I_14	Gossips or spreads rumors
I_20	Is disruptive
I_22	Breaks rules

Response Categories 0 = Never/Almost Never, 2 = Sometimes, 3 = Often, 4 = Very Often

Appendix D

Sensitivity Analysis: Hierarchical Multiple Regression

	Model	*b*	*SE*	*β*	*p*
1	Achievement	.01	<.01	.71	<.001
2	Achievement	.01	<.01	.57	<.001
	Effort	.41	.01	.33	<.001
3	Achievement	.01	<.01	.57	<.001
	Effort	.41	.01	.33	<.001
	Attendance	-.06	.01	-.05	<.001
4	Achievement	.01	<.01	.57	<.001
	Effort	.41	.01	.33	<.001
	Attendance	-.06	.01	-.05	<.001
	Grade 5	-.05	.01	-.04	<.001
5	Achievement	.01	<.01	.56	<.001
	Effort	.39	.01	.31	<.001
	Attendance	-.06	.01	-.05	<.001
	Grade 5	-.05	.01	-.04	<.001
	Problem Behavior	-.05	.01	-.04	<.001
6	Achievement	.01	<.01	.57	<.001
	Effort	.37	.01	.30	<.001
	Attendance	-.06	.01	-.05	<.001
	Grade 5	-.05	.01	-.04	<.001
	Problem Behavior	-.05	.01	-.04	<.001
	Male	-.11	.01	-.09	<.001
7	Achievement	.01	<.01	.55	<.001
	Effort	.37	.01	.30	<.001
	Attendance	-.07	.01	-.06	<.001
	Grade 5	-.05	.01	-.04	<.001
	Problem Behavior	-.05	.01	-.04	<.001
	Male	-.11	.01	-.09	<.001
	White	.09	.01	.07	<.001
8	Achievement	.01	<.01	.53	<.001
	Effort	.37	.01	.30	<.001
	Attendance	-.06	.01	-.05	<.001
	Grade 5	-.05	.01	-.04	<.001
	Problem Behavior	-.05	.01	-.04	<.001
	Male	-.11	.01	-.09	<.001
	White	.05	.01	.04	<.001
	FARM	-.10	.01	-.08	<.001

Note. Hierarchical multiple regression. Dependent variable is overall GPA 2009, which is the mean of Reading, Math, Science, and Social Studies GPAs.

Appendix E

Classroom Aggregated Results from Multiple Regression

	Model	*b*	*SE*	*β*	*p*
1	Classroom Achievement	.09	.01	.46	<.001
	Classroom Effort	.37	.06	.26	<.001
	Classroom Attendance	.11	.10	.05	.27
2	Classroom Achievement	.05	.01	.28	<.001
	Classroom Effort	.34	.07	.24	<.001
	Classroom Attendance	.05	.10	.02	.63
	Classroom Problem Behavior	-.04	.07	-.03	.55
	Classroom Proportion Male	-.20	.13	-.07	.12
	Classroom Disadvantage	-.13	.02	-.29	<.001

Note. All variables are the student-level variables aggregated to the classroom level.

References

Alexander, K. L., Entwisle, D. R., & Dauber, S. L. (1993). First-grade classroom behavior: Its short- and long-term consequences for school performance. *Child Development*, *64*, 801-814.

Allen, J. (2005). Grades as valid measures of academic achievement of classroom learning. *The Clearing House, 78*, 218-224.

America's Best High Schools. (2011, June). *Newsweek.* Retrieved from http://www.thedailybeast.com/newsweek/features/2011/americas-best-high-schools.html

Austin, S., & McCann, R. (1992, April). "Here's another arbitrary grade for your collection": A statewide study of grading policies. Paper presented at the annual meeting of the American Education Research Association, San Francisco, CA.

Asparouhov, T., & Muthén, B. (2010). Computing the strictly positive Satorra-Bentler chi-square test in Mplus. *Mplus Web Notes Number 12*.

Berger, J., Yiu, H. L., Vaganek, M., Nelson, D., Gottfredson, G., Rosenfield, S, Gravois, T., Vu, P., Shanahan, K. B., & Hong, V. (2011). Teacher utilization of Instructional Consultation Teams. Unpublished manuscript, University of Maryland, College Park.

Beswick, J. F., Willms, J. D., & Sloat, E. A. (2005). A comparative study of teacher ratings of emergent literacy skills and student performance on a standardized measure. *Education*, *126*, 116-137.

Boado, H. C. (2011). Primary and secondary effects in the explanation of disadvantage in education: The children of immigrant families in France. *British Journal of Sociology of Education, 32*, 407- 430.

Bodovski, K. (2010). Parental practices and educational achievement: Social class, race, and *habitus*. *British Journal of Sociology of Education, 31*, 139-156.

Bonesrønning, H. (2004). Do the teachers' grading practices affect student achievement? *Education Economics*, *12*(2), 151-167.

Brookhart, S. (1993). Teachers' grading practices: Meaning and values. *Journal of Educational Measurement*, 30, 123-142.

Brookhart, S. (1994). Teacher's grading: Practice and theory. *Applied Measurement in Education, 7*, 279-301.

Bruckman, K. (2010). Influence of problem behavior and teacher tolerance on student grades. Masters thesis, University of Maryland, College Park.

Bub, K. L., McCartney, K., & Willet, J. B. (2007). Problem trajectories and first-grade cognitive ability and achievement skills: A latent growth curve analysis. *Journal of Educational Psychology, 99,* 653-670.

Bursuck, W., Polloway, E., Plante, L., Epstein, M., Jayanthi, M., & McConeghy, J. (1996). Report card grading and adaptations: A national survey of classroom practices. *Exceptional Children, 62*, 1-10.

Carlson, L. (2003). Beyond assessment to best grading practice: Practical guidelines. In J. Wall & G. Walz (Eds.) *Measuring Up: Assessment Issues for Teachers, Counselors, and Administrators* (pp. 507-515). ERIC Clearinghouse version. ED480379

Chi-square difference testing using the Satorra-Bentler scaled chi-square [Mplus website content]. Retreived from http://www.statmodel.com/chidiff.shtml

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. (2nd ed.). Mahwah, NJ: Erlbaum.

Conners, C. K. (2001). *Conners' rating scales - revised: Technical manual.* Toronto: Multi-Health Systems, Inc.

Crosby, E. G., & French, J. L. (2002). Psychometric data for teacher judgments regarding the learning behaviors of primary grade children. *Psychology in the Schools*, 39, 235-244.

Cross, L. H., & Frary, R. B. (1996, April). Hodgepodge grading: Endorsed by students and teachers alike. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Davis-Kean, P. & Sexton, H. (2009). Race differences in parental influences on child achievement: Multiple pathways to success. Merrill-Palmer Quarterly, 55, 285-318.

Duckworth, A., & Seligman, M. (2006). Self-discipline gives girls the edge: Gender in self-discipline, grades, and achievement test scores. *Journal of Educational Psychology, 96,* 198-208.

Enders, C. (2010). *Applied Missing Data Analysis.* New York, NY: Guilford Press.

Flynn, J. (1997). *Teacher Rating Scale.* LaCrosse, WI: LaCrosse Area Dyslexia Research Institute, Inc.

Frary, R., Cross, L., & Weber, L. (1992, April). Testing and grading practices and opinions in the nineties: 1890s or 1990s? Paper presented at the annual

meeting of the National Council on Measurement in Education, San Francisco, CA.

Friedman, S. J., & Frisbie, D. A. (1995) The influence of report cards on the validity of grades reported to parents. *Educational & Psychological Measurement*, 55, 5-26.

Finney, J. W., Humphreys, K., Kivlahan, D. R., & Harris, A. H. S. (2011). Why health care process performance measures can have different relationships to outcomes for patients and hospitals: Understanding the ecological fallacy. *American Journal of Public Health, 101*, 1635-1642.

Gottfredson, G. D. (1981). Schooling and delinquency. In S. E. Martin, L. B. Sechrest, & R. Redner (Eds.), *New directions in the rehabilitation of criminal offenders* (pp. 424-469). Washington, DC: National Academy Press.

Gottfredson, G. D., & Gottfredson, D. C. (1999). *Development and Applications of Theoretical Measures for Evaluating Drug and Delinquency Prevention Programs: Technical Manual for Research Editions of What About You? (WAY).* Marriottsville, MD: Gottfredson Associates: Authors.

Gottfredson, N. C. (2008). An empirical evaluation of the disaggregated effects of educational diversity in national sample of law schools. Masters thesis, University of North Carolina, Chapel Hill.

Gottfredson, N. C., Panter, A. T., Daye, C. E., Allen, W. F., & Wightman, L. G. (2009). The effects of diversity in a national sample of law students: Fitting multilevel latent variable models in data with categorical indicators. *Multivariate Behavioral Research, 44*, 305-331.

Greville, E. (2009). A rose by any other name: Grading and assessment. *Assessment Update: Progress, Trends and Practices in Higher Education, 21*, 1-2, 13.

Grewal, R., Cote, J. A., & Baumgartner, H. (2004). Multicollinearity and measurement error in structural equation models: Implications for theory testing. *Marketing Science, 23*, 519-529.

Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudeck, S. du Toit, & D. Sörbom (Eds.), *Structural Equation Modeling: Present and Future — A Festschrift in honor of Karl Jöreskog*. Lincolnwood, IL: Scientific Software International, Inc.

Hanushek, E., & Luque, J. (2003). Efficiency and equity in schools around the world. *Economics of Education Review, 22*, 481-502.

Heck, R., & Thomas, S. (2009). *An Introduction to Multilevel Modeling Techniques, Second Edition*. New York: Routledge.

Herman, M. (2009). The black-white-other achievement gap: Testing theories of academic performance among multiracial and monoracial adolescents. *Sociology of Education, 82*, 20-46.

Hinshaw, S. P. (1992). Externalizing behavior problems and academic underachievement in childhood and adolescence: Causal relationships and underlying mechanisms. *Psychological Bulletin, 111*, 127-155.

Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods & Research, 33*, 188-229.

Johnson, W., McGue, M., & Iacono, W. G. (2006). Genetic and environmental influences on academic achievement trajectories during adolescence. *Developmental Psychology, 42*, 514-532.

Kline, R. (2011). *Principles and Practice of Structural Equation Modeling, Third Edition.* New York: Guilford Press.

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.

Martinez, J., Stecher, B., & Borko, H. (2009). Classroom assessment practices, teacher judgments, and student achievement in mathematics: Evidence from the ECLS. *Educational Assessment, 14*, 78-102.

McMillan, J. H., Myran, S., & Workman, D. (2002). Elementary teachers' classroom assessment and grading practices. *The Journal of Educational Research, 95*, 203-213.

Monteforte, L. (2006). Aggregation bias in macro models: Does it matter for the euro area? *Economic Modeling, 24,* 236-261.

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28*, 338-354.

Muthén, L. K., & Muthén, B. O. (1998-2010). *Mplus User's Guide*. Sixth Edition. Los Angeles, CA: Muthén & Muthén.

Novotny, J. S. (2011). Academic resilience: Academic success as possible compensatory mechanism of experienced adversities and various life disadvantages. *The New Educational Review, 23* 91-101.

Ouimet, M. (2000). Aggregation bias in ecological research: How social disorganization and criminal opportunities shape the spatial distribution of juvenile delinquency in Montreal. *Canadian Journal of Criminology, 42*, 135-156.

Polloway, E., Epstein, M., Bursuck, W., Roderique, T., McConeghy, J., & Jayanthi, M. (1994). Classroom grading: A national survey of policies. *Remedial and Special Education, 15*, 162-170.

Pomerantz, E., Altermatt, E., & Saxon, J. (2002). Making the grade but feeling distressed: Gender differences in academic performance and internal distress. *Journal of Educational Psychology, 94,* 396-404.

Psychological Corporation. (2002). *Wechsler individual achievement test: Second edition.* San Antonio, TX: Psychological Corporation.

Randall, J., & Engelhard, G. (2009). Differences between teachers' grading practices in elementary and middle schools. *The Journal of Educational Research, 102*, 175-185.

Randall, J., & Engelhard, G. (2010). Examining the grading practices of teachers. *Teaching and Teacher Education, 26*, 1372-1380.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*, 581-592.

Satorra, A. & Bentler, P. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66*, 507-514.

Stone, B. J. (1994). Group ability test versus teachers ratings for predicting achievement. *Psychological Reports*, 75, 1487-1490.

Strein, W., & Meshbesher, N. (2006). Grades and grading. In G. Bear and K. Minke (Eds.). *Children's Needs: Psychological Perspectives III*. Bethesda, MD: National Association of School Psychologists.

Shanahan, K. B., Vu., P., Rosenfield, S., Gravois, T., Vaganek, M., Berger, J., & Gottfredson, G. D. (2011). The effects of Instructional Consultation Teams on student outcomes. Unpublished manuscript, University of Maryland, College Park, Maryland.

Vaught, S. E. & Castagno, A. E. (2008). "I don't think I'm racist": Critical Race Theory, teacher attitudes, and structural racism. *Race, Ethnicity and Education, 11*, 95-113.

Virginia Standards of Learning Assessments. (2009). Technical Report: 2008-2009 Administration Cycle. Retrieved from http://www.doe.virginia.gov/testing/test_administration/technical_reports/sol_technical_report_2008-09_administration_cycle.pdf

Vu, P., Shanahan, K. B., Koehler, J., Rosenfield, S., Gravois, T., Berger, J., Vaganek, M., Kaiser, L., Nelson, D., & Gottfredson, G. D. (2011). Experimental evaluation of the effects of Instructional Consultation Teams on teacher efficacy, instructional practices, collaboration and job satisfaction. Unpublished manuscript, University of Maryland, College Park, MD.

Werthamer-Larsson, L., Kellam, S.G., & Wheeler, L. (1991). Effect of first-grade classroom environment on child shy behavior, aggressive behavior, and concentration problems. *American Journal of Community Psychology*, *19*, 585-602.