Why do students get good grades, or bad ones? The influence of the teacher, class, school, and student

Elaine M. Allensworth and Stuart Luppescu
University of Chicago Consortium on School Research
1313 East 60th Street
Chicago, IL 60637

Email: elainea@uchicago.edu
Email: lupp@uchicago.edu
Office: 773-702-3364
Fax: 773-702-2010

Please contact Elaine Allensworth for all correspondence.

Acknowledgements: We are grateful to our colleagues at the UChicago Consortium on Chicago School Research for their advice, critique and support on this work, particularly Shanette Porter, Bronwyn McDaniel, Alex Seeskin, Christopher Young, Lauren Sartain, Jenny Nagaoka, Jessica Tansey, Steffani Dornan, and Jessica Puller. We would like to thank the Bill and Melinda Gates Foundation for their funding, and the Consortium Investor Council for their broad support of work at the UChicago Consortium.

This is a working paper. Working papers are preliminary versions that are shared in a timely manner, with the aim of contributing to ongoing conversations in research and practice. They have not undergone the Consortium's full internal review process, nor have they received external peer review. Views expressed in this paper do not necessarily reflect those of the UChicago Consortium or the University of Chicago. Any errors are the authors' own.

Abstract

High school course grades are a primary source of information about students' academic readiness, yet they are often viewed as inconsistent measures of student achievement—influenced by idiosyncratic practices across schools and teachers, systematic differences in course content and structure, compositional effects of peers, and student demographics. Prior research has not quantified the extent of this variation relative to the influence of students' academic skills and effort, or examined multiple sources of influence simultaneously. This study employed cross-classified random-effects models with a dataset of 2.1 million grade records from 125,223 students and 11,000 teachers at 118 schools to identify sources of variation in students' grades. Grades varied based on which teacher students had for a given class and the conditions of the class (course subject, classroom peer achievement, time of day, class size and term). There were systematic differences in course grades by race and gender, even among students taking the same classes under the same conditions with the same test scores and attendance in the course. However, students' effort and skills, measured through attendance and test scores, dwarfed other sources of variation. Within- and between-student variation in attendance across classes also explained a substantial portion of variation by teacher, school and course conditions. Rather than finding large unexplained differences in grades based on which school a student attended, or which teacher they had, we found observable factors systematically explained differences in the grades that students received, particularly in students' aggregate grade point averages.

High school course grades are extremely important for students and their later outcomes, affecting college admissions, scholarships and credit accumulation toward high school graduation. Yet, grading practices vary from teacher to teacher, course to course and school to school. There is variation across teachers in how they grade assignments, and across classes in the material that is taught. There is also evidence that structural elements of classes—such as the class size, the composition of students, and the time in which it is offered—influence the grades that students receive. While many factors may influence grades, it is not clear to what extent each of them make a difference and potentially influence the comparability of grades across students.

Understanding the sources, and the extent of variation in students' grades is crucial for better using grades in educational decision-making. Grades are a primary source of information in early warning indicator and college readiness indicator systems (Borsato, Nagaoka, & Foley, 2013; Bowers, 2009; Balfanz, Herzog & MacIver, 2007). They are used to make decisions about needs for academic supports, and access to programs, schools, and colleges. Grade point averages are highly predictive of high school and college graduation, providing the best source of information about whether students will succeed in higher levels of school (Author, 2007; Bowers, 2010; Bowen, Chingos, & McPherson, 2009; Geiser & Santelices, 2007; Hiss & Franks, 2014; Hoffman & Lowitzki, 2005; Rothstein, 2004). Yet, a number of studies have raised concerns about variability in the meaning of GPAs, even suggesting that a "B" average at one school could represent the same skills as a "D" average at others (Godfrey, 2011; U.S. Department of Education, 1994; Woodruff & Ziomek, 2004). As school practitioners use grades for data-based decisions, information about the sources and extent of variation in grades could provide guidance on how confident people can be about students' underlying achievement based on their course grades, help practitioners understand why a student might receive a poor grade in a particular class, and be aware of course situations when students are more likely to need extra support to earn a good grade.

This study examines the degree of variation in students' high school course grades, and the sources of that variation, examining the factors that are associated with higher and lower grades as students take different classes under different conditions. Prior studies often have been based on samples of students or teachers in particular subjects (e.g, history), or have examined teachers' assessments of student work separate from their classroom practice, or used proxies for students grades (e.g., self-reported grades). This study uses the universe of actual core-course grades (2.1 million grade records) received over four years in a diverse array of 118 high schools. The methodology used in this study considerably extends the work that has been done in the past by using information about the clustering of students within classrooms, as well as within-student differences in grades to parse out teacher and classroom effects. It also contributes new information by concurrently examining the influence of a large array of classroom conditions on the grades that students receive, rather than studying them as isolated factors, including the period in the day the class is taken, the term (fall, spring, summer), class size, content (e.g., algebra, calculus, U.S. history), achievement level of peers, and students' achievement relative to their classroom peers.

Prior Literature on Sources of Variation in Course Grades

Teachers use different formulas for calculating grades, based on their own criteria around student performance, effort, and growth, on a potentially wide-ranging array of tasks, (e.g., project work, class discussion, worksheets, presentations, research papers, teacher-developed assessments, and textbook-based tests), and incorporate teacher judgment about individual students and strategies for motivating student effort (Bowers, 2011; Brookhart, 1993; Cross & Frary, 1996; Farkas, Sheehan, Grobe, & Shuan, 1990; Kelly, 2008; Willingham, Pollack, & Lewis, 2002). A number of studies have found variation in teachers' grading criteria and judgment, even when grading the same assignment (see Brookhart et al., 2016, for a review). The size of the differences in many cases is fairly small—equivalent to about 5 points on a 100 point scale, but, when considered along with the variability in course assignments and expectations, it introduces the potential for considerable random variation in the grades that students receive for similar effort and skills.

In general, educators, policymakers, and parents tend to view the role of schools as developing broad competencies in students, not just the skills that are measured on standardized tests (Bowers, 2011; Nagaoka, Farrington, Ehrlich, & Heath, 2015; Willingham et al., 2002), and grades reflect this broad emphasis. They incorporate a number of factors other than core content knowledge and academic skills that are believed to be important to prepare students for college and career, including effort (attendance, study habits), reliable assignment completion, class participation, time management, help-seeking behavior, metacognitive strategies, and social skills. This array of factors which matter for students' grades and educational attainment, but are not measured on tests, have been characterized in a number of ways, including noncognitive skills (Farrington et al., 2012), 21st Century Skills (National Research Council, 2013), and School Success Factors (Bowers, 2011). While grades provide a fuller picture of students' performance than standardized tests, the broad range of factors on which they are based introduces the possibility that there may be random variation in terms of what the grades represent, based on teachers' idiosyncratic decisions about course expectations and grading.

At the same time, it is possible that teachers' expectations and grading practices are not so different that they account for a substantial proportion of the variation in students' grades. The general expectations across classrooms may be sufficiently similar—attending every day, turning in all assignments, studying, meeting academic standards that are prescribed by the state and professional organizations, and using textbooks and curricular resources with the same general content, even if the specific tasks that students are asked to perform are different. It is also possible that students' GPAs could be stable indicators of students' achievement, even if grades in individual courses are strongly influenced by the particular teacher. Because students take many different courses in high school, any one grade will only contribute a small amount to their total GPA, and students will likely take a variety of classes with a variety of teachers. The variation in the grades students receive because of differences among teachers and course conditions might be small when grades are averaged into a GPA.

There also may be structural factors that influence the grades students receive in systematic ways—not all of the variation in students' grades across teachers and schools may be idiosyncratic and random. To the extent that the variation in grades are based on observable factors, these differences could be taken into account when assessing students' grades.

Frogpond (achievement composition) effects. One structural source of variation in grades could come from differences in the composition of students in a particular class. A number of studies have discerned "frogpond" effects (Attewell, 2001), where students with similar academic performance and efforts receive lower grades if they are in a class of high-achieving students than if they are in a class of low-achieving students (Farkas et al., 1990;

Kelly, 2008; Author, 2013), or if they enroll in schools with higher-achieving vs. lower-achieving peers (Author, 2016; Attewell, 2001; Barrow, Sartain, & de la Torre, 2016). These differences are often attributed to relativistic grading practices, where teachers evaluate students' performance compared to other students in the class. However, these patterns would also occur if teachers adjust their instruction to match the general skill level of the class, introducing more challenging material if their students have mastered basic content and slowing the pace of instruction if the class seems to need more time.

Intentionally challenging courses. It is generally acknowledged that it is harder to earn high grades if students take courses that have a reputation for being academically challenging, or are intentionally designed to be difficult. Sadler and Tai (2007) found that Honors and AP high school science courses predicted performance in subsequent college courses that were equivalent to 0.5 GPA points higher for Honors classes and a full GPA point higher for AP classes, compared to students in regular science classes. These courses may be designed to be more difficult, although frogpond effects may also contribute to students receiving lower grades in these classes than others, since it is difficult to parse out the effects of class composition from those of class content.

Other course-specific structural differences. Students' grades could also differ systematically by other features of classes, in ways that may be intentional, but often are not. Students receive lower grades in advanced science and math courses than in other subjects (Bassiri & Schulz, 2003). The time of day in which a course is taken affects students' attendance, effort, and performance (Wahistrom 2002; Randler & Frech 2009). Student effort on homework completion may wane when the weather gets warm in the spring, causing their grades to drop in the second semester. Class size may also affect students' achievement (Hedges, Laine, &

Greenwald,1994; Krueger, 2003). Differences across classes in time of day, time of year, or class size have nothing to do with teachers' expectations or grading practices, but they introduce variation into grades that may seem arbitrary if it is not understood.

Societal inequalities and stereotypes. Societal stereotypes may influence grades and also people's perceptions of grades. Sometimes arguments that grades are not equivalent seem to incorporate suggestions of racial bias—in both directions. Based on the frogpond effect, people sometimes assume that GPAs are not equivalent for students of different race and ethnic groups, or different economic backgrounds, because differences in average school achievement levels across groups. At the same time, theories of perceptual bias and self-fulfilling prophecies suggest that stereotypes in the broader society can influence teachers' perceptions of students' achievement, or affect students' interactions with teachers and behaviors in class, so that it may be harder for particular groups to earn high grades (e.g., Gilliam, Maupin, Reyes, Accavitti, & Shic, 2016; Williams, 1976). Stereotypes in the larger society can influence students' self-perceptions, attitudes toward learning, and academic performance through stereotype threat, (Steele & Aronson, 1995; Walton & Cohen, 2007), and teachers' judgements of students' work habits and skills influence their grades (Farkas et al., 1990).

Some studies that control for test scores and measures of effort show that Asian students tend to receive higher grades than other students, while Black students tend to receive lower grades, and that boys tend to receive lower grades than girls who have similar test scores and attendance (Farkas et al., 1990; Author, 2007). A number of other studies do not show clear differences in teachers' grading assessments by student socio-economic status (SES) (e.g., Leiter & Brown, 1985; Willlams, 1976). However, there are systematic differences in the types of courses and schools that students enroll based on race, ethnicity, gender and SES--it is not known to what extent grades differ among students taking similar courses under similar conditions.

Limitations of Prior Research

Research suggests there are a number of sources of influence on students' grades, but it has not quantified the degree to which these sources of influence are large, relative to the influence of students' effort and skills. Most of the studies have examined one or two factors at a time, when those relationships could be influenced by other factors not included in their analysis. Some studies have reported inconsistencies in grades based on a mismatch with test scores (Godfrey, 2011; U.S. Department of Education, 1994; Woodruff & Ziomek, 2004), but they used students' self-reported grades, rather than grades from transcripts, even though they are less precise (Zwick & Himmelfarb, 2011). These studies did not show variation across schools, and did not account for factors other than test scores when comparing grades.

It is also not known whether differences in the grades given by different teachers, in different schools, are mostly a result of random differences among them (e.g., subjectivity or idiosyncratic grading practices), versus systematic factors (e.g., class size and subject), versus differences in student effort under different contexts or with different teachers. Some teachers in some schools may be better able to motivate students to put in effort and learn, so that their students earn higher grades than with other teachers in other schools. Finally, research showing differences in students' grades by their background characteristics has not accounted for the many differences that exist in terms of the types of classes students take, the conditions of those classes, whether they have different kinds of teachers, or show different amounts of academic effort or skills.

Therefore, we ask:

1) To what extent do students receive different grades based on:
    a. Their teacher for a particular class?
    b. The school the student attends?
    c. The student's background characteristics?
    d. Specific conditions of the class?

2) How much of the variation in grades by school, teacher, student background, and class conditions can be attributed to students' academic skills and efforts, as measured by test scores and attendance?

3) What are the factors that account for systematic differences in grades across classes, teachers, and schools?

Data

This study makes use of an extensive Chicago Public Schools (CPS) dataset on students' grades in every core course in high school (English/Language Arts, Math, Science and Social Science), over multiple years, as well as their attendance in those courses, high school test scores (reading, English, math and science), and background characteristics, using more than 2.1 million grade records from 125,223 students and 10,327 teachers at 118 schools. Each grade that a student receives is linked to that student by a student ID number and to a teacher by a teacher ID number. Students can be grouped with other students in the same class by the course number, period, term, and teacher ID, allowing us to study compositional influences of classes.

The analyses are based on the population of students and teachers in Chicago public high schools, using semester course grades from all high school courses in core subjects (math, English, science, and social science) from the 2003-04 school year through the 2006-07 school year. All students who enrolled in a CPS high school at any time in that period were included in the analyses, including those who transferred in or out during the period, and those who eventually dropped out of school. However, the analyses were restricted to those who attended CPS in the eighth grade, so that we had measures of their prior achievement. Some students and teachers had grade records at more than one school (13.6 percent of students or 17,000 students; 2.4 percent of teachers); these improved the estimation of school effects above and beyond controlling for the characteristics of students and classes, since we could see differences in the grades the same student received across different schools. There is a wide variety of high schools in Chicago, ranging from extremely high-achieving selective schools that consistently rank among the top-performing schools in the country, heterogeneous schools, and very poor-performing schools with low graduation rates.

Tables A.1-A.3 in the Appendix provide descriptive statistics on each of the variables used in the study, including the frequencies of course grades, and the average grades associated with each independent variable. Unweighted grades of F through A, coded on a 5-point scale, 0 through 4, are the primary outcome. They are treated as numeric rather than ordered categories in the analysis; the differences between grade steps (e.g., F to D vs. B to A) were not sufficiently different to warrant a multinomial logit model, given the increase in complexity both for computation and interpretation. We did conduct an ordered category analysis with a subset of the data (a 25 percent random sample) which showed that the spacing between thresholds was fairly uniform, so we felt comfortable coding grades as continuous variables and interpreting the coefficients as the difference between each letter grade. Future studies may consider examining differences using ordinal outcomes to discern whether the patterns here differ at high versus low ends of the grading scale.

The modal grade was a C, with 24 percent of cases, while the average grade was 1.88 (see Tables A.1 and A.2 in the appendix for more information). In general, there are considerable differences in grades based on students' gender, race, prior test scores, and grade level (see Appendix Table A.3).

The analysis took into consideration a variety of classroom characteristics associated with each grade observation, including type of course (e.g., calculus, English I), course period, term (fall, spring, summer), and three measures of class composition (described below). Type of class is defined in students' transcript files, based on course title and course number. There is general consistency across schools in the numbering of core courses, particularly those courses that are used to fulfill graduation requirements. However, the specific curriculum (text book, supplemental materials) that are used for the same course may vary by school. Courses were included from four general areas—English/language arts, math, science, and social science/history. The list of included courses can be seen in Table 2 in the results section.

There are multiple measures of classroom composition. Class size was based on the number of students in the class where the grade was received. The average prior achievement level of classroom peers was calculated from those students' average math and English scores, using the eighth-grade achievement measure described below. We used student achievement from the eighth grade as a constant measure of achievement prior to high school for students in different grade levels. We included indicators for whether a student was taking a class with much higher- or lower-achieving peers than themselves when they earned that grade, using a variable coded 1 if the student was at least 0.5 standard deviations above the average entering achievement level of the class, and another variable coded 1 if the student was at least 0.5 standard deviations below the average entering achievement level of the class, and 0 otherwise for both variables. Standard deviations were derived from the population of students so that the meaning is constant regardless of the heterogeneity of the class. By including indictors of both the average achievement level of the class, and the student's relative position in the class (whether much higher or lower achieving than typical), we could parse out differences in grades that were due to general changes in expectations based on classroom averages from differences that were based more on relativistic grading practices.

At the student level, we included information on students' backgrounds, including incoming achievement, gender, race/ethnicity, and whether the student started high school older than age 14. Prior achievement was based on students' eighth-grade latent math and reading scores. The latent scores were calculated based on the Illinois Standards Achievement Test (ISAT) math and reading scores available for students from the third- to eighth-grade year. This is a more precise indicator of prior achievement than a single eighth-grade test score because it takes into consideration the student's entire test score trajectory. We also included measures of SES in the student's residential census block group based on Census data; an index of poverty, based on the percentage of adult males unemployed and the percentage of families with incomes below the poverty line, and a measure of social status that includes the mean level of education of adults and the percentage of employed persons who work as managers or professionals. Census block groups are defined by the U.S. Census Bureau and in Chicago typically represent a city block. These provided a more precise measure of students' SES than a dichotomous indicator of qualification for free or reduced-price lunch. The student's grade level at the time they took the class was a background variable included at the observation level, since their grade level is not constant.

We used course absence and standardized test scores to measure students' academic effort and skills. Effort involves more than showing up for class, and standardized test scores are incomplete measures of the academic content and skills that students learn in all their classes. Thus, this study likely provides an underestimate of the degree to which grades represent real differences in academic effort and skills across students, which is a limitation.

Each grade record includes the number of days the student was absent in the class. Absence has a positive skew, but analyses produced similar results whether or not it was transformed to have a normal distribution. Therefore, we show models where class absence is entered in days, so that it is easy to interpret the coefficients. We also included a squared term to allow for a nonlinearity that exists in the relationship between absences and grades.

Academic skills were measured through the EPAS exams. During the years of this study, students took the EXPLORE at the beginning of ninth grade, the PLAN at the beginning of tenth grade and again at the beginning of eleventh grade, and the ACT at the end of the eleventh grade; each included subject tests in math, science, English, and reading. The content of the tests was not strongly aligned with the curriculum in any one class, so they were used to measure students' general level of academic achievement during their high school years. We combined the scores from all of the EPAS tests into one overall measure of high school achievement, standardizing the scores by subject, grade level, and test (EXPLORE, PLAN, ACT) and averaged them together. In this way, students who left school before they took all of the tests had values that represented their scores for the tests that were taken while they were still in school. We did not expect this measure to explain differences in a given student's grades from course-to-course, but we included it to potentially explain differences in students' overall GPAs.

Analytic Approach

We leveraged the fact that students receive grades from many different teachers to discern teacher effects net of student effects—finding the degree to which students systematically received higher or lower grades with particular teachers compared to the grades those same students received from their other teachers. Those estimates were further adjusted for the conditions under which the grade was received (e.g., the type of course, the classroom peers). The logic of the approach is shown in Figure 1. Each grade was simultaneously nested within a student (who received grades for multiple classes) and a teacher (who gave out grades to hundreds of students). That grade was also tied to a particular course, with a particular subject, time of day, class composition. In Figure 1, a "B" grade is given to Ana (the student) by Mr. Jones (the teacher) in a specific class (3rd period Algebra):

- Student random main effects represent the underlying skills and traits that each student brings to their courses that leads them to get a particular grade. They are calculated as the average grade a student earns across all of their classes, adjusted for characteristics of the classes in which the grades were earned, and the random effects of the teachers that assigned each of their grades (whether those teachers tend to give higher or lower grades than typical).

  For example, in Figure 1, let's assume Ana comes to the class with strong skills and work habits that she tends to demonstrate across all of her classes, so she usually gets high grades. Her "main effect" might show she is a "B+/A-" student that would be expected to have a 3.5 average in typical classes with average teachers. There are unique conditions in any class that affect her grade, so sometimes she gets an A, and other times a B. In her class with Mr. Jones she gets a “B,” perhaps because he tends to give out lower grades than typical (with a negative teacher main effect).

- Teachers’ random main effects are calculated as the average grade they give out, adjusted for the random main effects of the students who received those grades (e.g., whether the students in their classes tend to get high/low grades in their other classes with other teachers), and for the characteristics of the class in which they gave each grade to each student (e.g., subject, period, size). Differences in teacher effects could result because they are easy or hard graders, or because they are particularly effective or ineffective at motivating and teaching students.

  For example, in Figure 1, Mr. Jones tends to give out low grades, relative to the grades his students get from other teachers (giving Ana a “B” when she often gets As), but the specific grade he gives to any given student will depend on the student and the class conditions. He gives a “B” to Ana because she shows strong skills and effort, but other students get “Cs” and “Ds.”

- The conditions of the class (fixed effects) add a further element of variation, so that the same student and teacher combinations could produce different grades if the course is offered first period vs. third period, or is a calculus class instead of an algebra class.

Because Ana's class is third period, attendance might be especially high, so that Mr. Jones gives out fewer low grades in that class than in others he teaches.

Statistical Models

We ran a series of increasingly complex models in which we were primarily interested in the variance components, and how they changed across models as more variables were added, to answer RQs 1 and 2. Model 1 simply nests grades within students to discern the extent to which students received different grades among their classes (within-student variance), and the degree to which students' average grades (GPAs) were different from each other (the between-student variance). The combined model is:

$$Y_{ij} = \theta_0 + r_{0j} + e_{ij} \qquad (1)$$

where

Y_{ij} is the grade observation *i* for student *j*

θ_0 is the average grade (GPA) across all students

r_{oj} is the individual random effect of student *j*; its variance is the between-student variance in GPAs.

e_{ij} is the individual random effect of student *j* in class *i*; its variance is the within-student, between-class variance in course grades.

We then used a cross-classified analysis for all subsequent models, which nests each course grade simultaneously within the student that received it and the teacher that assigned it, using the R function lmer in the lme4 package (see Bates, Maechler, Bolker and Walker, 2015). The variance in teacher effects shows the degree to which individual teachers systematically gave higher or lower grades than other teachers gave to the same students when those students were in the other teachers' classes. The change within-student variance with the addition of teacher random effects shows the degree to which differences in students' grades from class-to-class are systematically associated with particular teachers. Following this, we added a school variance component to discern the degree to which there are systematic differences in grades across schools (Model 2b):

$$Y_{ijkl} = \theta_0 + r_{oooj} + v_{000k} + o_{oool} + e_{ijkl} \qquad \text{(2b)}$$

Where Y_{ijkl} is the grade observation *i* received by student *j* with teacher *k* in school *l*, θ_0 is the grand-mean average grade, and there are four random effects:

r_{000j} the random main effect of students

v_{000k} the random main effect of teachers

o_{000l} the random main effect of schools

e_{ijkl} the within-student residual.

We then included variables for students' prior test scores and backgrounds (Model 3). These models showed the degree to which students' characteristics prior to high school explained differences among students in the grades they received. The teacher and school variance components from these models showed the degree to which students with similar backgrounds and skills coming into high school received different grades, based on which school they attended or which teacher they had during high school:

$$Y_{ijkl} = \theta_0 + \boldsymbol{\pi}_{\boldsymbol{mj}}\boldsymbol{a}_{\boldsymbol{mj}} + r_{000j} + v_{000k} + o_{000l} + e_{ijkl} \qquad \text{(3)}$$

Where $\boldsymbol{a}_{\boldsymbol{mj}}$ represents m={1, ..., 24} indicators of prior achievement and background for student *j*.

The next models included classroom conditions to discern whether the variation in student, teacher, and school effects was reduced by specific conditions of the classes in which the grades were earned:

$$Y_{ijkl} = \theta_0 + \pi_{mj}a_{mj} + \beta_{ni}W_{ni} + \beta_{pi}W_{pi} + u_{000j} + v_{000k} + o_{000l} + e_{ijkl} \qquad (4)$$

Where W_{ni} represents n={1 to 24} indicators of class conditions for observation *i* and W_{ip} p={1 to 50} vectors of course subject indicators.

To address the second question, we incorporated variables representing the number of days the student was absent for the grade observation, and students' average scores on the EPAS tests. Models 5a and 5b include just course absence or EPAS, while Model 5c includes both:

$$Y_{ijkl} = \theta_0 + \pi_{mj}a_{mj} + \pi_{12j}EPAS_j + \pi_{13i}Days\ absent_i +$$

$$\beta_{ni}W_{ni} + \beta_{pi}W_{pi} + u_{000j} + v_{000k} + o_{000l} + e_{ijkl} \qquad (5c)$$

For RQ 3, we examined the coefficients associated with contextual factors, from Models 4 and 5c, rather than the variance components.

Results

The first row of Table 1 shows the variance components from an unconditional model that simply nests grades within students (Model 1). Because there are no other variables, the variance of student effects (0.96) is the between-student variance in core course GPAs. The variance components are not easily interpretable, but they can be transformed into standard deviation units by taking their square root, and then used to approximate the distribution of grades net of the variables in the model. In this case, the standard deviation is 0.98, which corresponds to a four standard deviation range (95 percent range) of 3.9, which is about the range of GPAs from 0.0 to 4.0. The within-student residual variation shows how much students' grades vary from their overall grade point average. There is almost as much variation in course grades within students, across the classes they take (0.89), as there is in average grades across students (0.95). This suggests that most students get a variety of grades across their classes.

The next series of models incorporate teacher effects, then school effects, followed by student's incoming skills and backgrounds, and finally classroom characteristics. The unconditional models (Model 1, 2a and 2b) provide a base from which to compare changes in the variance components as new variables are added. If a variance component shrinks when new variables are added, it suggests that those variables explain a portion of that variance.

RQ 1: To what extent do students' grades vary based on their teacher, their school, or the specific conditions of a class?

Teacher effects. Model 2a cross-nests grades simultaneously within teachers and students. The teacher random main effects show the degree to which teachers give out grades that are higher or lower, on average, than the grades their students receive with other teachers. There is considerable variation across teachers (variance component of 0.38). Accounting for teacher effects decreases the within-student variance from 0.87 to 0.70. Thus, teacher effects account for about one-fifth of the within-student variation in grades (why the same student gets different grades in different classes). In contrast, the variance component for student main effects is almost the same with the inclusion of teacher effects as without them (0.95 versus 0.96). Students take a variety of classes with a variety of teachers, and differences across teachers seem to even out, explaining very little of the differences in students' overall GPAs.

School effects. The next model (Model 2b) incorporates school effects. The variance in school effects is much smaller (0.10) than the variance across students (0.94) or teachers (0.35). Differences across schools also explain little of the variance in grades across teachers (the teacher main effect decreases from 0.38 to 0.35) or students.

Students' prior test scores and backgrounds. Adding variables for students' prior test scores in Model 3a reduces the student-level variance component from 0.94 to 0.81, explaining about 13 percent of the variance in students' GPAs. The school-level and teacher-level variance components increase slightly once prior test scores are included. Differences in grades across teachers and schools are somewhat larger when we just compare students with similar eighth-grade test scores than when we don't take into account students' prior test scores. This is consistent with the frogpond or compositional effects; high-achieving students are more likely to attend schools with negative effects on grades.

Model 3b brings in other background characteristics, such as students' race, gender, SES, age when starting high school, and grade level. These background characteristics account for about 9 percent of the total variance in students' GPAs, reducing the student level variance from 0.81 to 0.72. The coefficients are discussed in detail with RQ3. Controlling for background characteristics slightly reduces the teacher-level and school-level variance.

Altogether, Models 3a and 3b explain one-quarter of the variance in student main effects (decreasing from 0.94 to 0.72). Thus, while prior test scores and background characteristics explain a quarter of the variation in GPAs across students, substantial variation remains unexplained. This means that there are large differences in high school grades among students with the same incoming test scores and background characteristics, attending the same schools.

Characteristics of classes. Model 4 brings classroom characteristics into the equations, including the course subject (e.g., U.S. History, biology), the average achievement level of the class, the student's achievement relative to peers in the class, class period, term, level (e.g., Honors, AP), and the class size. Taking into account classroom characteristics reduces the variation across teachers by one-third (from 0.34 to 0.23). Thus, a substantial portion of the variation in grades across teachers can be attributed to characteristics of the classes those teachers teach. The degree to which particular classroom characteristics are related to students' grades is discussed further below, with RQ3.

RQ 2: How much of the variation in grades by school, teacher, and class conditions can be attributed to students' academic skills and efforts (standardized test scores and attendance)?

The next series of models show whether the differences in grades observed across students, the teacher and school differences in grades, can be attributed to students' high school test (EPAS) scores and the number of days they were absent in each course. The addition of EPAS scores (Model 5a) only slightly reduces the variance of student main effects (from 0.77 to 0.75). The prior models already control for students' incoming test scores, and high school standardized test scores may not capture substantial differences in academic skills beyond what is already captured by these earlier tests.

Adding course attendance to the model substantially reduces the size of all of the variance components (Model 5b), compared to Model 4, dwarfing the changes that occurred with the inclusion of the variables in prior models. About half of the remaining variance in students' main effects is explained by course attendance; the variance component declines from 0.77 to 0.37. High school grades are much more strongly related to whether students come to class regularly than to their race, gender, SES, or prior test performance. The final model (5c) incorporates both EPAS scores and attendance and has the smallest remaining unexplained variation in student main effects, although it is similar to the prior model.

The within-student residual variance also shrinks when course attendance is added to the model (from 0.67 to 0.59). Thus, one reason students get better grades in some classes than others is that they are more likely to attend particular classes. One-fifth of the remaining teacher variance is explained by student attendance (declining from 0.23 to 0.18), suggesting that students are more likely to attend classes taught by some teachers than by others, and this explains some of the differences in the grades teachers give. There still is considerable variance in teacher effects (0.18), but about half of the differences among teachers are explained by course characteristics and student attendance. There is very little remaining school-level variance after adding student attendance to the models (0.05 in Model 5c). Differences in GPAs across schools are largely explained by differences in student attendance and differences in course characteristics, including average student achievement levels.

To get a sense of how a student's grade point average might differ based on which school they attend, Figure 2 displays the school-level variance components from Models 3a and 5c transformed into standard deviations in GPA units. The top two bars show school-level differences in grades for a student who would receive a 2.0 GPA at a typical school with typical teachers. Although the overall average GPA is 1.88, we use 2.0 as the example for ease of presentation, and because C is the modal grade. The very top bar compares students who start high school with similar characteristics; a student who would earn a 2.0 GPA at a typical school would be likely to end up with a 2.4 GPA if he attended a school where students tend to get higher grades than typical (one standard deviation above the mean), and a GPA of 1.6 if at a school where students tend to get lower grades than typical (one standard deviation below the mean). These GPAs are considerably different, but still are generally within the range of what would be considered a C student. The differences are larger among schools at the extremes—at a school at the 98th percentile the student would likely end up with a GPA of 2.7 (B-), while he would end up with a GPA of 1.3 (D+) at a school at the 2nd percentile.

As shown in Table 1, about half of the differences between schools in students' grades are a result of differences in students' attendance, and there are also differences based on course characteristics (class peer ability levels and type of courses taken). The second bar of Figure 2 shows the degree to which students earn different grades across schools, comparing students who not only enter high school with similar skills and backgrounds, but also have similar attendance while in high school, and who take similar kinds of classes. With this comparison, school effects are modest. Students at schools that are one standard deviation above or below the mean have GPAs that are 0.2 points higher or lower than at a typical school, respectively. Even at the extreme schools that give out the highest or lowest grades to similar students with similar attendance, the 2.0 student would end up with a GPA in the C range (between 1.6 and 2.4).

The teacher-level variance is much larger than the school-level variance. As shown in the third bar in Figure 2 (the first bar representing teacher effects), a student who would get a 2.0 with an average teacher (a solid C), would be likely get a 2.6 (B-) with a teacher that had a high positive effect, and a 1.4 (D+) with a teacher with a large negative effect. About a quarter of the differences between teachers are explained by the characteristics of the classes they teach. As shown in the next bar on Figure 2, comparing teachers who teach similar classes, there are smaller differences. One further factor to consider is that students are more likely to attend classes taught by some teachers than others. The final bar in Figure 2 compares grades among teachers of similar students, in similar classes, with similar days absent in those classes. The student who would get a solid C with an average teacher would likely have a grade of 2.4 (C+) if she had a teacher for the same class with a large positive effect (one standard deviation above the mean), and a 1.6 (C-) grade from a teacher with a large negative effect (one standard deviation below the mean), if the student was present the same number of days in both classes. At the extremes, the student would end up with a grade that was one letter grade higher or lower than she would receive with a typical teacher.

RQ3: What are the factors that account for systematic differences in grades across classes, teachers and schools?

Tables 2a-2c provide the coefficients from models 5a and 5c. They are divided by student characteristics, course characteristics other than subject, and course subject so that the tables are not too long. However, the variables from all of the tables were included together in Models 5a and 5c. The coefficients from Model 5a show the relationship of each variable in the model with students' grades, net of the other variables in the model. Because the sample size is so large, almost all coefficients are significantly different from zero, although some are modest in size; standard errors are included in the table so readers can gauge the confidence level for each.

Some of the relationships of student or course characteristics with grades are influenced by student attendance. As shown at the top of Table 2a, each day of absence in a class is associated with a decline in the grade of about 0.07 grade points (adjusted for the squared term). Missing 10 days (two school weeks, about a day a month) would be associated with a decline of 0.6 grade points (taking the squared term into consideration). The difference between the coefficients from the two models provides an indication of the degree to which the relationship of that student or course characteristic with grades may be attributable to student attendance.

Student characteristics. The first set of coefficients shows the relationships of student background characteristics with grades, net of the other variables in the models. On average, Black and Latino students get lower grades than White students (by 0.26 and 0.12 GPA points, respectively), while Asian students get higher grades (by 0.26 points). These differences exist even though the models control for students' test scores and economic status, are compared relative to students in the same types of classes and in the same schools, and control for students' skills relative to their classroom peers. The differences by race and ethnicity are largely not explained by student attendance in the class; attendance explains about one-third of the difference in grades received by Asian American students (shrinking from 0.484 to 0.311), about 14 percent of the difference in grades received by Black students (shrinking from -0.257 to -0.219), and none of the difference in grades for Latino students.

Boys receive grades that are about half of a grade point lower than girls with similar test scores taking similar classes under similar conditions. The gender difference also remains strong after controlling for students' course attendance (0.31). Thus, there are substantial racial, ethnic, and gender-based differences in grades that are not explained by attendance, test scores, the courses in which students enroll, class composition, students' skills relative to their peers, or which teacher they have for those courses. These differences are additive, so a Black male student would have grades that are 0.64 points lower, on average, than a White female student with the same attendance and test scores who takes the same types of classes in the same school.

There are modest-to-moderate differences in grades based on the economic status and poverty level of students' neighborhoods, moderate differences in grades for students entering high school at different ages, and large differences in grades based on students' incoming test scores. Course attendance explains a substantial part of the difference in grades based on economic factors, as well as the lower grades received by students who enter high school old-for-grade. About one-third of the relationship between prior test scores and grades is related to differences in attendance among students with different test scores. Higher-achieving students earn higher grades in part because they attend class more often.

If we do not use nested models to study grades, it appears that grades are lowest in the ninth grade (see Appendix Table A.3). However, that is because students with low course grades tend to drop out of school at higher grade levels. The nested models show that students' grades are actually slightly lower in tenth grade and beyond. The effect reverses when attendance is included. This suggests that students tend to miss classes more frequently at older grades, and their grades would be higher in later grades if they attended class more often.

Course characteristics. Table 2b shows the relationships of class structure variables with grades. The average achievement of classroom peers is negatively associated with course grades. Students tend to receive lower grades when they are in classes with peers who have higher levels of prior achievement, compared to the peer achievement levels in their other classes (lower by 0.173 points for each standard deviation increase in peer achievement). This relationship exists, controlling for the subject that is being taught and whether it is an Honors or AP class. It suggests that teachers adjust their expectations to the incoming skills of the students in the class, so that it is harder to get good grades in classes with more high-achieving students. There is only slight evidence for frogpond effects, or "grading on a curve," in that students get higher grades when they have exceptional incoming skills relative to their classroom peers (by 0.011 points), and lower grades when they have much lower skills (by 0.010 points). The frogpond effects are much smaller than the effect of overall class achievement level—even students with the highest incoming skills get lower grades in classes with high-achieving peers than in their classes with low-achieving peers.

The negative relationship between classroom peer ability level and course grade becomes even stronger when attendance is controlled (comparing Model 5c to Model 5a), suggesting that students attend classes with high-achieving peers more than their classes with lower-achieving peers, but have a harder time earning high grades in those classes. This is consistent with the hypothesis that teachers increase the rigor and expectations for classes with higher-achieving students, and vice-versa. In a similar vein, students receive lower grades in AP and Honors classes than they do in their regular classes. These coefficients also are larger controlling for attendance (Model 5c vs Model 5a)—students get lower grades in Honors and AP classes even though they are more likely to attend these classes than their other classes.

Course period, term, and class size. There are systematic differences in students' grades based on which period they take a class. Grades are highest during third period; coefficients for all the other periods are negative, with grades being particularly low during first period (0.193 points lower, on average). Once we control for attendance, all of the period effects become much smaller, with several becoming non-significant. The effect of class size (all relative to the smallest category; fewer than 15 students) shows a fairly consistent pattern wherein students get lower grades in larger classes. These differences largely remain after taking attendance into account. Grades are slightly lower in spring than in fall term, and this is completely explained by attendance—the coefficient even flips to be positive once attendance is controlled. Summer classes have higher grades, but students can miss no more than one class session in the summer to receive credit for the class, which is a strong incentive to attend. Attendance explains much of the difference in average grades during the summer.

Course subject. Table 2c shows the differences in grades associated with course subject, all relative to algebra I. Grades tend to be lower for the more advanced courses, even controlling for the average incoming achievement level of students in the classes, and their grade level. Thus, grades are not lower in these classes just because they are taken by more high-achieving students at older grades. Students' grades tend to be particularly low in advanced math and science classes (pre-calculus, calculus, advanced life science), and tend to be higher in remedial classes and in language arts electives (e.g., journalism, creative writing, drama). A number of advanced science and math courses show much larger negative effects when we do not control for AP status (calculus, chemistry II, physics II), suggesting they are often taken as AP classes (alternative model results available from authors). With some of the courses where students tend to receive lower grades, the effects are attenuated by attendance. This suggests grades are lower in these classes partly because students miss these classes more often. Advanced math courses (geometry, algebra II, pre-calculus, and calculus), and some science courses (physics, chemistry) fall into this group. In other cases, the coefficients are larger after controlling for attendance. For example, grades are higher in computer science and drama than other classes, and the coefficients are even larger when attendance is controlled. This suggests that students tend to get higher grades in those classes even though they also tend to miss those classes more than is typical.

The types of classes in which students enroll influence students' grades; this can be seen in these coefficients, and in the reduction of variance that occurred in the teacher and school effects, and in the residual (within-student) variation, when class characteristics were added to the models in Table 1. Because students tend to take a wide variety of classes, some that are associated with higher grades and others with lower grades, many of these differences offset each other, such that adding course characteristics shows only a small influence on the variance in student main effects. However, for a student that takes a large proportion of coursework that tends to be more or less difficult for earning high grades, these could be meaningful differences, and they do explain some of the differences in GPAs across schools.

We can use the coefficients in Tables 2b and 2c to calculate the difference that course characteristics make, on average, for students with very demanding schedules compared to those with undemanding schedules. For example, taking only Honors classes with high-achieving peers (1 standard deviation above the mean) would likely reduce a student’s core course GPA by 0.25 of a GPA point (the class average effect of -0.173 plus the Honors class effect of -0.075). If one-quarter of a student’s classes were AP classes, instead of Honors classes, still with high-achieving peers, then the GPA would be lower by 0.33 points. If four of those classes in the prior schedule were pre-calculus, calculus, physics, and chemistry II, the GPA would be lower by 0.39 points relative to an undemanding schedule. Thus, a very demanding schedule could make a difference of about 0.40 GPA points, relative to an undemanding one.

Discussion

People often express the belief that grades are not objective indicators of student achievement, given that there are different expectations and grading practices among teachers and schools. We find that there are differences among teachers and schools in the grades that students with similar backgrounds receive, taking classes under the same conditions; however, the differences are not as large as often believed, and much of the variation can be explained by observable factors. Variation by teacher is considerably larger than variation by school. The grade a student receives in any given class may not be a good representation of her overall level of achievement, and could be as much as a letter grade different (0.8 GPA points) from what she would receive from a more typical teacher, under the same conditions. Because students take many different classes with many different teachers, teacher differences tend to average out in terms of their contribution to students' overall GPAs, having little overall influence on them. It is possible that some students might systematically have those teachers who give out particularly high or low grades relative to other teachers in their school teaching similar courses, but this seems to be rare.

Neither differences in grades by teachers nor across schools accounts for more than a small proportion of the differences in students' overall GPAs, particularly when comparing students taking similar courses under similar conditions, with similar test scores and attendance. Teacher-level variance is one-third the size of student-level variance, and half of that teacher variance is explained by the characteristics of the courses teachers teach, and the level of student attendance in their courses. School-level variance is almost completely explained by observable factors. This suggests some degree of consistency in assigning grades among education professionals; the standards for grades across schools may not be as arbitrary as is often believed. Rather than finding large unexplained differences in grades based on which school a student attends, or which teacher they have, we find there are observable factors that systematically explain most of the differences in the grades that students receive in different types of schools, and with different teachers.

Attendance and tested skills. By far, the factors that are most strongly associated with differences in students' GPAs are their course attendance and tested skills. These measures of academic effort and skills explain much more of the differences in students' average grades than the teachers they have, the courses that they take, or the school they attend, even though attendance and standardized test scores are very limited measures of students' academic effort and skills. Attendance also explains one-half of the variance between schools--students get better grades at some schools because they are more likely to show up for their classes at those schools. Students are also more likely to show up in some classes and with some teachers than others, and they tend to receive better grades in those classes with those teachers than in their other classes.

The fact that attendance so strongly predicts students' grades may raise concerns that students receive grades just for seat time. More likely, attendance is a proxy for general work effort and learning, such that students who show up more often may participate more in class, put more time into studying and getting assignments done, and produce better quality work. It also may be that attendance is more crucial for learning than people realize. If students' miss class, they not only receive less instruction but they can fall behind in assignments and in understanding course material—the effects may not be obvious since students never experience the counterfactual of attending a class they missed. Attendance explains over 2.5 times more variance in students' GPAs than their test scores, and even seems to account for one-third of the relationship between test scores and GPAs (the coefficient associated with incoming test scores shrinks by one-third when attendance is added to the model). Even in classes taken by high-achieving students, such as chemistry, physics, and calculus, students seem to get lower grades partly because they are less likely to show up for those classes than their other classes. We cannot say that there is a causal relationship between attendance and course grades based on this analysis—there may be external factors that affect both students' attendance and their grades, or students may simply show up to their classes more often when they are earning high grades.

However, the very large size of this relationship suggests that improving student attendance in high school has more potential leverage for improving their grades than strategies that are focused mostly on test scores. This is an area where we would strongly suggest more research to discern a causal effect.

Course types and structures. The context in which a student takes a course makes a difference for their grade, so that assessments of students' achievement based on grades should take the course subject and conditions into account. Students' grades are influenced by the time of day, the term, the class size, and their grade level when they took the class, which contribute to the fact that any one grade may not be representative of a student's overall achievement. It is possible that schools could systematically design classes in ways that could support students to get better grades by paying attention to these factors, particularly for students who are at risk of poor performance. For example, schools might not schedule core courses first period for students at risk of failure. Overall, these conditions even out when averaged, so that they do not have much influence on overall GPAs.

The conditions that are more likely to influence students' GPAs are the course subjects and peer-achievement levels of their classes. A student that took a demanding schedule, with Honors and AP classes and advanced-level science and math courses, with high-achieving peers, would be likely to have a GPA that was about 0.40 GPA points lower than he would have with an undemanding schedule. These differences could matter for whether a student would be chosen for a highly-competitive program, such as an extremely competitive college that uses very small differences in GPAs to make decisions about students. It would make sense to consider coursework and school achievement level when making high-stakes decisions based on grades, which often occurs with decisions such as college admissions.

Student background. One concerning finding from these analyses is the large differences in grades that exist based on students' race and gender, even when comparing students with similar backgrounds, test scores, attendance, and coursework, in similar classes with the same teachers in the same schools. That these differences exist above and beyond all of these other factors would seem to be a critical area of study, given concerns about the underrepresentation of racial-ethnic minority and male students in college. Often, it is difficult to separate out differences in students' grades from the non-random sorting of students into classes and schools, but that is not the case here.

Future research could use models similar to those used here to examine interactions of teacher and class characteristics with race and gender, or enter aspects of classroom instruction, bring in discipline records, or test hypotheses about stereotypes, stereotype threat, differences in discipline policies, outside influences, supportive practices of teachers, to determine under what conditions these differences are reduced. Another question is whether GPAs underpredict future performance for boys or racial-minorities, since they receive lower grades net of the measures included here representing skills and effort. If so, it suggests issues around grading bias. If not, it indicates that the factors that lead them to lower grades in their high school classes are also relevant for later outcomes, and critical to understand to increase equity in educational attainment.

Future Research

There are many questions that arise as a result of this study, and a number of ways in which the analysis could be improved to be more comprehensive. It would be invaluable to have additional measures of students' skills and effort in their classes other than attendance and standardized test scores. It is striking that two such very limited measures of effort and skills could explain so much of the differences in students' grades. Because those measures are limited, we are likely over-estimating the degree to which there are idiosyncratic teacher and school effects on grades, as well as the degree of residual (within-student) variation in grades. Perhaps, as districts develop more advanced electronic gradebook systems, such analyses will be possible in the future. It would also be worthwhile to add teacher information to the models, to learn whether there are specific background, training, or attitudinal characteristics that are related to students' grades and to look for interactions between those teacher characteristics and the characteristics of the students themselves. Some students may benefit more from particular types of teaching or teachers; this might also help explain some of the residual (within-student) variation. Finally, this analysis is based on the Chicago Public Schools and the results may not be generalizable to districts that serve very different student populations. While there are a wide range of high schools in CPS, including very selective schools and very low-achieving schools, not all results may be the same in other places with different types of schools serving different student populations. Future research might replicate the findings in other places that have extensive longitudinal datasets.

Research Ethics

This research was conducted with Institutional Review Board approval.

References

Author, (2007)

Author, (2013)

Attewell, P. (2001). The winner-take-all high school: Organizational adaptations to educational stratification. *Sociology of Education 74*(4), 267-295.

Balfanz, R., Herzog, L., & MacIver, D.J. (2007). Preventing student disengagement and keeping students on the graduation path in urban middle-grades schools: Early identification and effective interventions. *Educational Psychologist*, *42*(4), 223-235.

Barrow, L., Sartain, L., & de la Torre. (2016). *The role of selective high schools in equalizing educational outcomes: Heterogeneous effects by neighborhood socioeconomic status*. Chicago, IL: University of Chicago Consortium on School Research.

Bassiri, D., & Schulz, E. M. (2003). Constructing a universal scale of high school course difficulty. *Journal of Educational Measurement*, *40*(2), 147-161.

Bates, Douglas, Maechler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software 67,* 1.

Borsato, G.N., Nagaoka, J., & Foley, E. (2014). College readiness indicator systems framework. *Voices in Urban Education, 38,* 28-35.

Bowen, W.G., Chingos, M.M., & McPherson, M.S. (2009). *Crossing the finish line: Completing college at America's public universities*. Princeton, NJ: Princeton University Press.

Bowers, A.J. (2009). Reconsidering grades as data for decision making: More than just academic knowledge. *Journal of Educational Administration*, *47*(5), 609-629.

Bowers, A.J. (2010). Grades and graduation: A longitudinal risk perspective to identify student dropouts. *The Journal of Educational Research*, *103*(3), 191-207.

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*(3), 141-159.

Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., & Welsh, M. E. (2016). A century of grading research: Meaning and value in the most common educational measure. *Review of Educational Research, 86*(4), 803-848.

Brookhart, S.M. (1993). Teachers' grading practices: Meaning and values. *Journal of Educational Measurement*, *30*(2), 123-142.

Cross, L.H., & Frary, R.B. (1999). Hodgepodge grading: Endorsed by students and teachers alike. *Applied measurement in Education, 12*(1), 53-72.

Farkas, G., Sheehan, D., & Grobe, R.P. (1990). Coursework mastery and school success: Gender, ethnicity, and poverty groups within an urban school district. *American Educational Research Journal*, *27*(4), 807-827.

Farkas, G., Grobe, R.P., Sheehan, D., & Shuan, Y. (1990). Cultural resources and school success: Gender, ethnicity, and poverty groups within an urban school district. *American Sociological Review*, *55*(1), 127-142.

Geiser, S. & Santelices, V. (2007). *Validity of high school grades in predicting student success beyond the freshman year: High-school record versus standardized tests as indicators of fouryear college outcomes*. Berkeley, CA: University of Berkeley Center for Studies in Higher Education

Gilliam, W.S., Maupin, A.N., Reyes, C.R., Accavitti, M., & Shic, F. (2016). *Do early educators' implicit biases regarding sex and race relate to behavior expectations and recommendations of preschool expulsions and suspensions?* New Haven, CT: Yale Child Study Center.

Godfrey, K. E. (2011). Investigating grade inflation and non-equivalence. Research Report 2011-2, College Board.

Hedges, L.V., Laine, R., & Greenwald, R. (1994). Does money matter? A meta-analysis of studies of the effects of differential school inputs on student outcomes. *Education Researcher, 23*(3), 5–14.

Hiss, W.C., & Franks, V.W. *Defining promise: Optional standardized testing policies in American college and university admissions.* Arlington, VA: National Association for College Admission Counseling (NACAC).

Hoffman, J.L., & Lowitzki, K.E. (2005). Predicting college success with high school grades and test scores: Limitations for minority students. *The Review of Higher Education*, *28*(4), 455-474.

Leiter, J., & Brown, J. S. (1985). Determinants of elementary school grading. *Sociology of Education*, 166-180.

Kelly, S. (2008). What types of students' effort are rewarded with high marks?. *Sociology of Education*, *81*(1), 32-52.

Kelly, S.P. (2008). Social class and tracking within schools. In L. Weis (Ed.), *The way class works: Readings on school, family and the economy* (pp. 210-224). New York, NY: Taylor & Francis Group.

Krueger, A.B. (2003). Economic considerations and class size. *The Economic Journal*, *113*(485), F34-F63.

National Research Council. (2013). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. Washington, DC: National Academies Press.

Nagaoka, J., Farrington, C.A., Ehrlich, S.B., Heath, R.D., Johnson, D.W., Dickson, S., Turner, A.C., Mayo, A., & Hayes, K. (2015). *Foundations for young adult success: A developmental framework*. Chicago, IL: University of Chicago Consortium on Chicago School Research.

Pattison, E., Grodsky, E., & Muller, C. (2013). Is the sky falling? Grade inflation and the signaling power of grades. *Educational Researcher*, *42*(5), 259-265.

Randler, C., & Frech, D. Young people's time-of-day preferences affect their school performance. *Journal of Youth Studies, 12*(6), 653-667.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage.

Sadler, P.M., & Tai, R.H. (2007). Weighting for recognition: Accounting for advanced placement and honors courses when calculating high school grade point average. *NASSP Bulletin, 91*(1), 5-32.

Steele, C.M. & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, 69(5), 797-811.

U.S. Department of Education, Office of Educational Research and Improvement. 1994. *What Do Student Grades Mean? Differences across Schools.* Office of Research Report 94-3401. Washington, D.C.

Wahistrom, K. (2002). Changing times: findings from the first longitudinal study of later high school start times. *NASSP Bulletin*, *86*(633), 3-21.

Walton, G.M. & Cohen, G.L. (2007). A question of belonging: Race, social fit, and achievement. *Journal of Personality and Social Psychology,* 92(1), 82-96.

Williams, T. (1976). Teacher prophecies and the inheritance of inequality. *Sociology of Education*, 223-236.

Willingham, W.W., Pollack, J.M., & Lewis, C. (2002). Grades and test scores: Accounting for observed differences. *Journal of Educational Measurement, 39*(1), 1-37.

Woodruff, D. J., & Ziomek, R. L. (2004). ? Differential Grading Standards among High Schools. ACT Research Report Series, 2004-02. *ACT Inc*.

Zwick, R. (2013). Disentangling the role of high school grades, SAT® scores, and SES in predicting college achievement. *ETS Research Report Series, 2013*(1), 1-20.

Zwick, R., & Himelfarb, I. (2011). The effect of high school socioeconomic status on the predictive validity of SAT scores and high school grade-point average. *Journal of Educational Measurement*, *48*(2), 101-121.

Tables

Table 1. Variance components from models predicting course grades

MODEL		Student variance (variation in GPAs $var(u_{000j})$)	Within-student residual (variance in individual student grades $var(r_{ijkl})$)	Teacher variance (var(v_{000k}))	School variance ($var(o_{000l})$)
1	Course grades nested within students only, no covariates	0.958	0.873		
2a	Cross-nested model with students and teachers, no covariates	0.946	0.698	0.377	
2b	Cross-nested model with school effects, no covariates	0.943	0.697	0.346	0.095
3a	Adding students' prior test scores	0.805	0.694	0.367	0.140
3b	Adding student background[1]	0.723	0.691	0.336	0.130
4	Adding course type, level, average classroom achievement, term, period, class size category [2]	0.768	0.673	0.234	0.109
5a	Adding average EPAS scores, without attendance[4]	0.754	0.673	0.234	0.115
5b	Adding student attendance[3] in the course	0.367	0.592	0.183	0.042
5c	With average EPAS scores and attendance	0.354	0.592	0.182	0.048

Variance components are bolded when the difference compared to the prior model is greater than 0.02.

[1] Student background covariates entered at the student level include race/ethnicity, gender, old-for-grade, SES, and concentration of poverty in the student's residential census block group. In addition, students' grade level is included at the observation (course grade) level.

[2] Classroom conditions are entered at the observation level.

[3] Student attendance is entered at the observation as the number of days the student is absent in the course from which he received the grade represented by that observation. The number of days squared is also included.

[4] EPAS system tests measure students' general skills in reading, English, math, and science in grades 9-11.

Table 2a: Coefficients from the full models in grade point units: Student variables

		Model 5a *Without Absences*		Model 5c *With Absences*	
		Estimate	Std. Error	Estimate	Std. Error
Intercept		**2.297**	0.033	**2.538**	0.022
Course absences (in days)				**-0.070**	0.000
Course absences squared				**0.001**	0.000
EPAS score – average of all high school EPAS tests		**0.340**	0.007	**0.313**	0.005
Student background variables at the student level					
Latent Eighth-Grade Achievement (test scores)		**0.255**	0.007	**0.166**	0.005
Student race and ethnicity (relative to White students)	Black	**-0.257**	0.011	**-0.219**	0.008
	Asian American Students	**0.484**	0.016	**0.311**	0.011
	Latino Students	**-0.124**	0.011	**-0.144**	0.007
Male (relative to female students)		**-0.466**	0.005	**-0.423**	0.004
Neighborhood SES		**0.014**	0.003	**0.006**	0.002
Neighborhood Poverty		**-0.055**	0.004	**-0.020**	0.003
Old-for-grade when entered high school		**-0.105**	0.003	**0.020**	0.003
Student variables at the observation level					
Grade when taking the class (relative to grade 9)	Grade 10	**-0.039**	0.003	**0.064**	0.003
	Grade 11	**-0.011**	0.003	**0.164**	0.003
	Grade 12	**-0.010**	0.004	**0.263**	0.004

These models control for teacher and school effects, as well as the characteristics of classes taken by students.

Table 2b: Coefficients from the full models: Course characteristics other than course subject

		Model 5a *Without Absences*		Model 5c *With Absences*	
		Estimate	Std. Error	Estimate	Std. Error
Classroom Average Math Achievement		**-0.173**	0.003	**-0.184**	0.003
Achievement compared to classroom peers	Higher: 0.25 sd > average	**0.011**	0.003	**-0.008**	0.002
	Lower: 0.25 sd < average	**-0.010**	0.003	**0.015**	0.002
Course level, (relative to regular)	AP	**-0.421**	0.006	**-0.436**	0.005
	Honors	**-0.075**	0.003	**-0.088**	0.003
Class period (relative to third period)	1	**-0.193**	0.002	**-0.030**	0.002
	2	**-0.050**	0.002	**-0.003**	0.002
	4	**-0.012**	0.003	**-0.002**	0.002
	5	**-0.039**	0.003	**-0.009**	0.002
	6	**-0.063**	0.003	**-0.018**	0.002
	7	**-0.087**	0.002	**-0.025**	0.002
	8	**-0.100**	0.002	**-0.017**	0.002
	9	**-0.120**	0.003	**-0.021**	0.003
	Other (non-standard period)	**-0.060**	0.004	**0.017**	0.004
Class size (relative to not more than 15 students)	>15 and <= 20	**-0.111**	0.002	**-0.080**	0.002
	>20 and < =25	**-0.136**	0.003	**-0.111**	0.002
	>25 and < =30	**-0.152**	0.003	**-0.132**	0.003
	>30 and <= 35	**-0.160**	0.005	**-0.147**	0.005
	>35 and <= 40	**-0.128**	0.012	**-0.165**	0.011
	>= 40	**-0.209**	0.011	**-0.206**	0.010
Term (relative to fall)	Spring	**-0.071**	0.001	**0.053**	0.001
	Summer	**0.851**	0.009	**0.277**	0.008

These models control for teacher and school effects, the incoming characteristics of students (demographic characteristics and prior test scores), and the course subject.

Table 2c: Coefficients from the full models: Course subject

	Model 5a *Without Absences*		Model 5c *With Absences*	
Difference Relative to Algebra I:	Estimate	Std. Error	Estimate	Std. Error
Computer	**0.183**	0.032	**0.258**	0.030
Basic Math	0.308	0.215	**0.195**	0.202
Problem Solving	**0.132**	0.006	**0.100**	0.005
Workshop Math	**0.459**	0.033	**0.543**	0.031
Transition Math	**0.240**	0.030	**0.079**	0.028
Advanced Math Support	**0.249**	0.029	**0.271**	0.027
Applied Math	**0.012**	0.019	**0.060**	0.018
Geometry	**-0.148**	0.005	**-0.096**	0.005
Algebra II	**-0.265**	0.005	**-0.194**	0.005
Advanced Math	**-0.285**	0.008	**-0.189**	0.008
Pre-calculus	**-0.423**	0.010	**-0.357**	0.010
Calculus	**-0.185**	0.015	**-0.121**	0.014
Other science	**0.390**	0.032	**0.416**	0.030
Other Science	**0.156**	0.022	**0.217**	0.021
Basic Science	**0.059**	0.028	**0.081**	0.026
Earth Science	**0.058**	0.009	**0.096**	0.008
Secondary Physical Science	**-0.064**	0.038	**0.005**	0.035
Advanced Physical Science	**-0.004**	0.025	**0.059**	0.023
Physical Science	**0.103**	0.028	**0.178**	0.026
Basic Physical Science	**0.131**	0.113	**0.148**	0.105
Biology	**-0.017**	0.009	**0.011**	0.008
Elective Life Science	**-0.008**	0.016	**0.082**	0.015
Advanced Life Science	**-0.157**	0.016	**-0.089**	0.015
Physics	**-0.173**	0.012	**-0.089**	0.011
Physics II	**-0.101**	0.024	**-0.022**	0.023
Chemistry	**-0.129**	0.010	**-0.052**	0.009
Chemistry II	**-0.091**	0.021	**-0.020**	0.019
World Studies	**0.047**	0.008	**0.052**	0.008
US History	**-0.001**	0.008	**0.029**	0.008
History III	**-0.012**	0.009	**0.039**	0.008
Social Science III - Economics	**-0.005**	0.011	**0.038**	0.010
Social Science III - Psychology	**0.000**	0.014	**0.040**	0.013
Other social Science	**0.320**	0.015	**0.316**	0.014
Other Social Science III	**0.067**	0.009	**0.115**	0.008
English I	**0.098**	0.008	**0.093**	0.007
English II	**0.006**	0.008	**0.036**	0.007
English III	**-0.058**	0.008	**-0.006**	0.008
English IV	**0.000**	0.009	**0.044**	0.008
English Drama	**0.306**	0.011	**0.370**	0.010
AVID English	**0.041**	0.068	**0.001**	0.063
Other English	**0.048**	0.015	**0.116**	0.014
Basic English	**0.281**	0.027	**0.204**	0.026
Applied English	**0.211**	0.033	**0.237**	0.030
English Supplementary	**0.220**	0.009	**0.207**	0.008

English Lab	**0.179**	0.009	**0.203**	0.009
Other Literature	**0.094**	0.011	**0.129**	0.011
English ESL	**0.084**	0.054	**0.093**	0.050
Creative Writing	**0.118**	0.012	**0.186**	0.011
Journalism	**0.248**	0.013	**0.295**	0.012
Other English	**0.124**	0.010	**0.140**	0.009

These models control for teacher and school effects, students’ demographic characteristics and prior test scores, and course characteristics other than subject.

Figures

Figure 1. Sources of Variation in Students' Grades

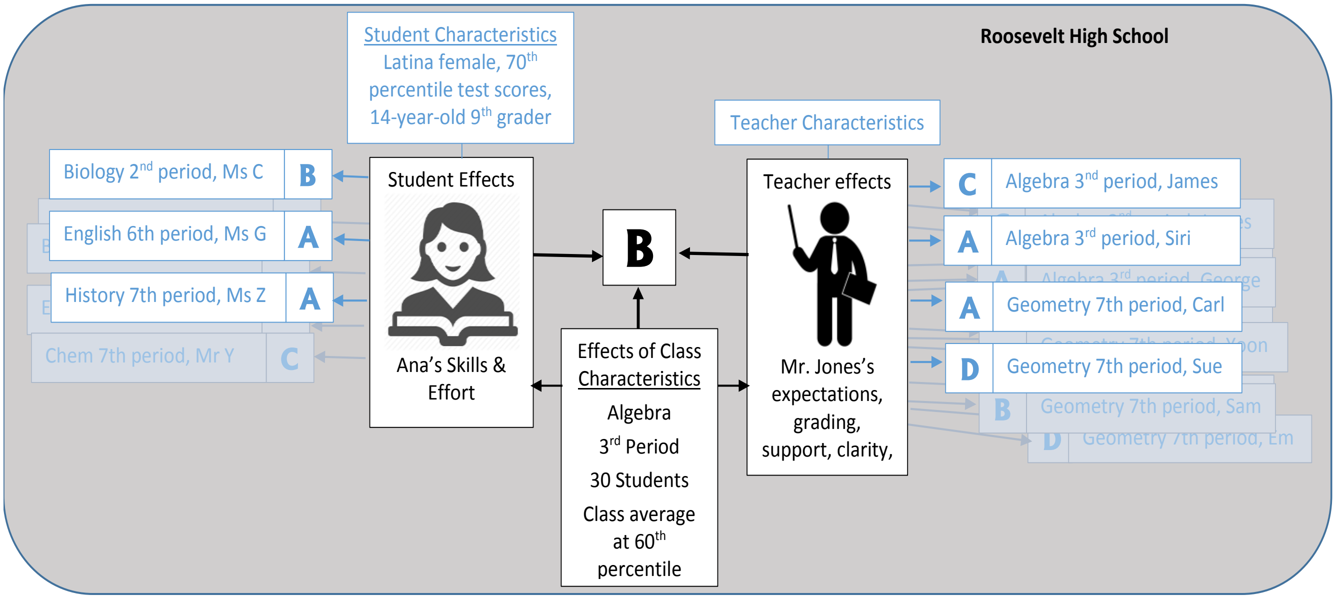
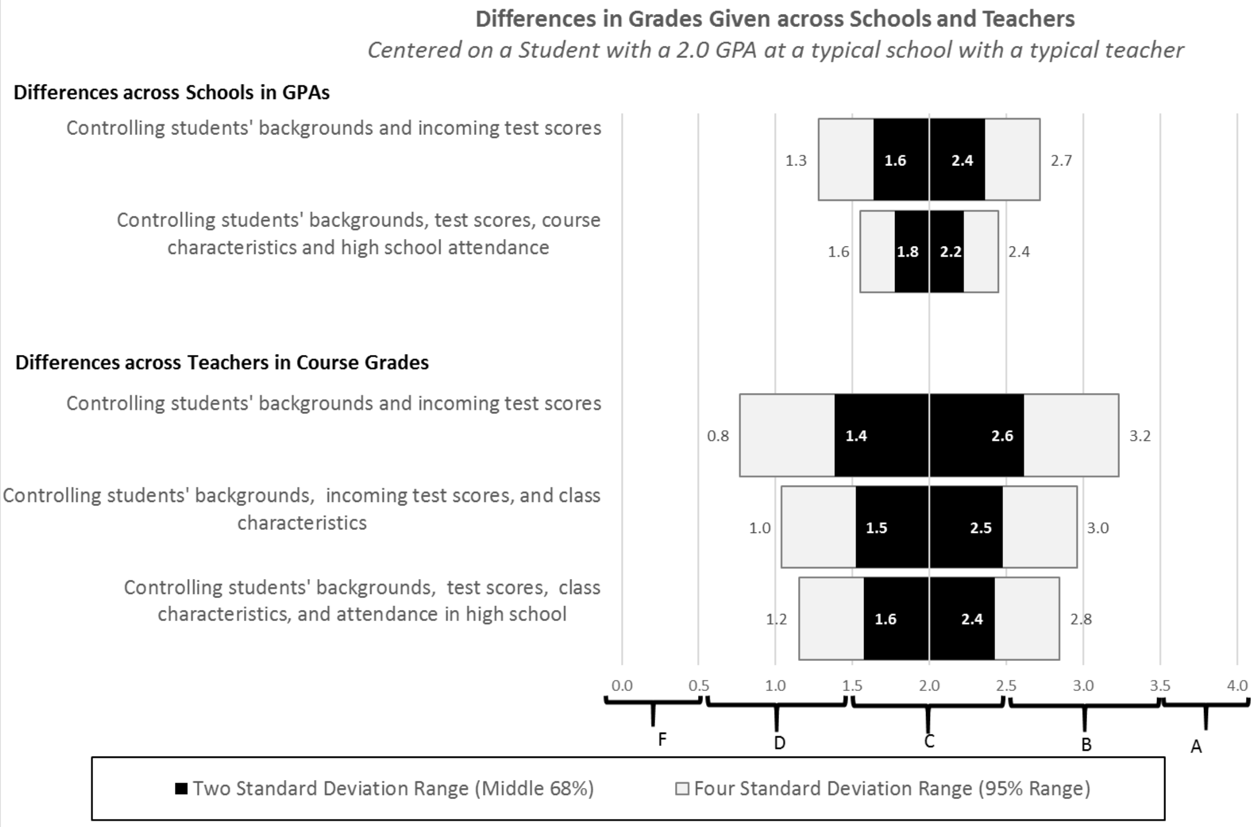


Figure 2.



Grade distribution ranges are calculated from the variance components in Table 2, taking the square root of the variance to get the standard deviation, and then multiplying by two or four to calculate the range of the student, teacher and school effects. Distributions are centered around students with a 2.0 GPA.

Appendix

Table A.1. Frequencies of categorical variables

	Frequency	%		Frequency	%
Course Grades			**Class Size**		
A	302,180	14.12	<=15	347,993	16.26
B	444,098	20.75	>15 and <=20	479,206	22.39
C	514,684	24.05	>20 and <=25	824,542	38.53
D	447,435	20.91	>25 and <=30	416,120	19.44
F	431,828	20.18	>30 and <=35	48,352	2.26
Achievement >.5 s.d. higher than class			>35 and <=40	7,419	0.35
0	1,739,609	81.28	>40	16,593	0.78
1	400,616	18.72	**Course Level**		
Achievement >.5 s.d. lower than class			AP	64,049	2.99
0	1,749,173	81.73	Honors	371,708	17.37
1	391,052	18.27	Regular	1,704,468	79.64
Class Period			**Course Term**		
1	259,299	12.12	Fall	1,175,575	54.93
2	285,149	13.32	Spring	938,562	43.85
3	268,007	12.52	Summer	26,088	1.22
4	225,711	10.55	**Old for grade starting high school**		
5	217,681	10.17	No	92,139	73.58
6	219,211	10.24	Yes	33,084	26.42
7	239,724	11.2	**Race/Ethnicity**		
8	251,515	11.75	Black	67,712	54.08
9	100,354	4.69	Latino	42,197	33.7
Other	73,574	3.44	White	10,950	8.75
Grade Level			Asian American	4,354	3.48
9	739,825	34.57	**Gender**		
10	598,772	27.98	Female	64,943	51.86
11	462,485	21.61	Male	60,280	48.14
12	339,143	15.85			

The unit of analysis is the grade observation. For example, 35% of the grade records were attached to students who were ninth graders at the time the grade was earned.

Table A.2. Descriptive statistics for continuous variables

	Mean	Std Dev
Course-Level Variables (n=2,140,225)		
Grades	1.88	1.33
Average classroom achievement (math & reading combined)	-0.01	1.01
Course absences	11.04	14.12
Student-Level Variables, all standardized (n=125,223)		
Incoming achievement (math& reading combined)	-0.04	0.98
Student residential average concentration of poverty	0.02	1.00
Student residential average social status	-0.01	0.99
Average EPAS scores	-0.03	1.00

Table A.3. Average grades by select student characteristics

		Mean	S.D.
Gender	Female	2.10	1.32
	Male	1.63	1.30
Race	Black	1.73	1.30
	Asian	2.76	1.20
	Latino	1.89	1.33
	White	2.30	1.33
Eighth-Grade test scores	Bottom quartile	1.45	1.25
	2nd quartile	1.65	1.29
	3rd quartile	1.97	1.31
	Top quartile	2.44	1.26
Grade level when taking the class	9th	1.66	1.34
	10th	1.81	1.32
	11th	2.03	1.30
	12th	2.27	1.27
Attendance in the class	0--2 absences	2.54	1.13
	2--6 absences	2.24	1.17
	6—14 absences	1.79	1.18
	14—99 absences	0.72	0.99