TITLE: Sources of variation in the connection between student grades and test scores.

Background, context, and purpose.

Grade inflation has been a persistent lament of educational researchers and commentators alike, with public criticism over the abundance of high grades and the lack of excellence needed to obtain them dating to at least the late 19th century (Kohn, 2002). Whether grade inflation is a problem needing to be solved, however, is contested. This debate is made more complicated because researchers have used the term grade inflation itself to encompass several distinct phenomena.

Researchers have most commonly studied grade inflation by looking at the extent to which the relationship between course grades and test outcomes has changed over time. For example, several studies have found that more recent cohorts of students receive better grades for the same tested performance as prior cohorts (Camara et al, 2003; Hurwitz & Lee, 2018). While most agree that the average student's GPA has increased over time, it is less certain whether more recent grades are now worse predictors of test performance (Pattison, Grodsky & Muller, 2013) or postsecondary outcomes (Brookhart et al., 2016). Researchers have also taken grade inflation to mean instances where grades are higher than expected given test results. For example, Gershenson (2018) found that 36% of North Carolina students who earned a B and about 70% of students earning a C in their Algebra I course were *not* deemed proficient on the state's end-of-course assessment.

Within both types of grade inflation, researchers have also given attention to how it varies by school: whether some schools' grades have risen or fallen faster than expected given test performance Hurwtiz & Lee (2018), or whether the same course grade could mean different things depending on the school in which it was earned (Godfrey, 2011). The extent which the types of student attending schools explain this variation has also been a focus. For example, Hurwtiz & Lee (2018) found schools that enroll wealthier students were more likely to have a larger increase in grades over time without a concomitant increase in test performance.

Though studies of grade inflation in some way always look at the connection between course grades and test scores, the aspect of this connection under study has not been consistent. This paper posits that the lack of a consensus on the meaning of grade inflation is because there are multiple dimensions to the connection between grades and test scores, each controllable by the schools and teachers bestowing grades, and each with a different implication for how parents, students, colleges, and the public should interpret the meaning of a grade. Specifically, we argue that the connection between grades and test scores can more precisely be characterized into three concepts: grade severity, grade alignment, and grade guarantee.

- **Grade severity** represents the extent to which earning better grades implies having learned more of the academic content in question. Systems that are more severe have bigger differences in tested results between students who earn the highest grades and students who earn lower grades.

- **Grade alignment** represents how well grades predict test scores. In a highly aligned system, we're able to obtain an accurate estimate for how a student will perform on the associated test based on the grade earned in class.

- **Grade guarantee** represents the likelihood that a student who earned a specific grade in a class has learned the class's content. Systems that have more grade guarantee are those where a higher proportion of students who, say, earn a B or higher demonstrate proficiency on the associated standardized assessment.

To illustrate these ideas, Figure 1 displays how students in a single class could be assigned grades in ways that highlight the differences between the three concepts. Importantly, all six examples below contain the same 30 students, with the exact same test scores (standardized against the state average). That is, the six examples illustrate that given a class of students, teachers (or systems) can make grading choices that demonstrate various levels of severity, alignment, and/or guarantee.

This study adds to the relatively thin existing evidence documenting grade severity, alignment, and guarantee. While many studies have relied on self-reported multi-year grade point averages and performance on college entrance examinations, this study uses students' actual student grades and state and national assessments directly linked to the course or grade-level. More importantly, this is the first study to estimate the sources of variation in these two concepts. By taking advantage of full course records, we can better understand the extent to which the connection between grades and test scores vary by classroom, teacher, and school.

Research questions.
RQ1: What is the grade distribution, severity, alignment, and guarantee for the average classroom?

RQ2: To what extent do grading distributions, grade severity, alignment, and guarantee vary between classrooms, teachers, and schools?

RQ3: Which student, classroom, and school characteristics are associated with grading distributions, grade severity, alignment, and guarantee?

Setting and population.
We use data from three large school districts, each serving over 10,000 K-12 students. Districts provided historical course grade and state test records for all students enrolled in the districts between the 2012-2013 and 2016-2017 school years. We constructed two analysis samples of students: First, we limited analyses to 6-8th grade math and ELA classes as these grade-levels provide end-of-year tests as well as teachers with multiple sections of the same course, which was desirable to estimate both classroom and teacher variation. Second, we analyzed all core subject Advanced Placement (AP) high school courses because of the direct connection between the course's content and the AP test. In all, our analyses include roughly 200,000 student-by-course combinations, 8,000 courses, and 2,000 teachers.

Research design and analysis.
Research Question 1
We use simple descriptive statistics to answer RQ1. For grade severity and guarantee, we identify the typical test performance for each course grade earned (e.g., C: 70-79, B: 80-89, A: 90-100). Using both standardized and binary test outcomes allows us to determine the probability a student with a given grade sufficiently met the expectations of the test (guarantee), and the difference in performances between students earning different course grades (severity). For grade alignment, we simply look at the correlation between test outcomes and course grades.

Research Question 2
For RQ2, we use a multi-level modeling framework to estimate variation in grade guarantee and grade alignment, with the mostly unconditional model:

$$Test\ Score_{icts} = \alpha_{0cts} + \alpha_{1cts}((Course\ Grade - 83)/10)_{icts} + e_{icts}$$

$$\alpha_{0cts} = \beta_{00ts} + r_{0cts}$$
$$\alpha_{1cts} = \beta_{10ts} + r_{1cts}$$

$$\beta_{00ts} = \pi_{000s} + u_{00ts}$$
$$\beta_{10ts} = \pi_{100s} + u_{10ts}$$

$$\pi_{000s} = \gamma_{0000} + v_{000s}$$
$$\pi_{100s} = \gamma_{1000} + v_{100s}$$

Where we model the test score (standardized against the full population of test takers of the same grade and subject or course in the same school year) of student *i* in classroom *c* of teacher *t* in school *s* as a function of that student's course grade (centered and re-scaled). This model provides point estimates for the typical test score for students earning a B (i.e., grade guarantee) in γ_{0000}, and the typical difference in test performance with a 10-point grade increase in γ_{1000} (i.e. grade severity). The *r* are the random "classroom effects", representing the deviation of classroom *c*'s grade guarantee and severity away from their teacher mean and *u* and *v* are the equivalents for teachers away from their school mean and schools away from the grand mean.

The Level-1 residual Is normally distributed, $e_{icts} \sim N(0,\ \sigma^2_{icts})$, where following Leckie et al. (2014) we can model the log of σ^2_{icts} as

$$\log(\sigma^2_{icts}) = \delta_{0000} + a_{0cts} + b_{00ts} + d_{000s}$$

Thus, δ_{0000} provides an overall estimate of grade alignment and *a, b,* and *d* provide estimates for how the extent to which a set of teachers' students' test scores depart from what was estimated based on the classroom, teacher, and school respectively.

We estimate the model using a fully Bayesian approach with Markov chain Monte Carlo (MCMC) methods and vague, flat, or minimally informative prior distributions for all parameters.

Research Question 3

Using the same model as RQ2, we add additional student, classroom, and school variables. This includes students' race/ethnicity, gender, FRPL status, IEP status, and prior achievement at the student level, and aggregated student characteristics at the classroom and school level. We also include controls for subject and academic year.

Selected findings.

We find a substantial disconnect between students' grades and test outcomes. Though grades were positively associated with test outcomes, in most districts 6th-8th grade students earning a B had less than a 25% probability of demonstrating grade-level performance on their state tests (Table 1).

We also found a meaningful proportion of variation in both grade guarantee and alignment between schools, teachers, *and* classes of the same teacher. Combined, the meaning of a student's grade can vary dramatically depending on the classroom. In all three districts, the estimated difference in test scores for the same course grade between the top and bottom quartile of classrooms was at least 1.25 standard deviations.

The grades of low-income students and students of color tended to demonstrate the least guarantee and alignment. Though recent research has demonstrated that affluent schools are inflating grades at faster rates (Gershenson, 2018, Hurwitz & Lee, 2018), it is traditionally disadvantaged students who currently receive grades that provide the least guarantee of test performance.

Conclusion.

Understanding what grades signal about student academic performance is timely. Recent research has highlighted the importance that parents give grades, who rely on them to know whether their child is achieving at grade-level, and who overwhelmingly agree that classroom grades provide a more accurate picture of achievement than state tests (Learning Heroes, 2017). This study suggests that parents' faith in grades as guarantees of learned content is not often well placed, and that what their child's grade actually signals depends heavily on which classroom door they walk into.

References

Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., Stevens, M. T. & 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.

Camara, W., Kimmel, E., Scheuneman, J., & Sawtell, E. A. (2003). *Whose grades are inflated?* (Research Report No. 2003-4). New York, NY: College Entrance Examination Board.

Kohn, A. (2002). The dangerous myth of grade inflation. The chronicle of higher education, 49(11), B7.

Gershenson, S. (2018). *Grade inflation in high schools (2005–2016).* Washington, DC. Thomas B. Fordham Institute.

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

Hurwitz, M. & Lee, J. (2018). Grade inflation and the role of standardized testing. In Buckley, J., Letukas, L., & Wildausky, B. (Eds.) *Measuring Success: Testing, grades, and the future of college admissions*. Baltimore, MD: Johns Hopkins Press.

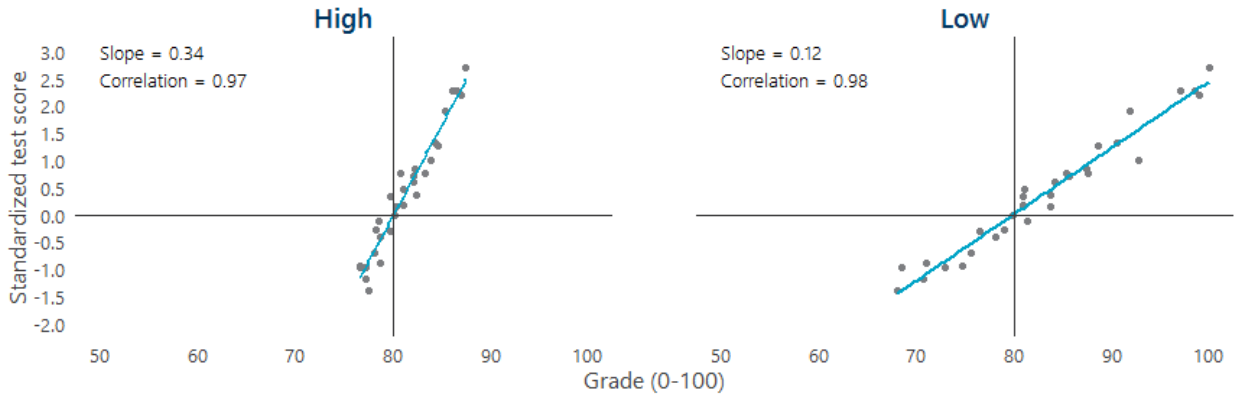
Learning Heroes (2017). *Parents 2017: Unleashing their power & potential* (Survey report: August 2017).

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.

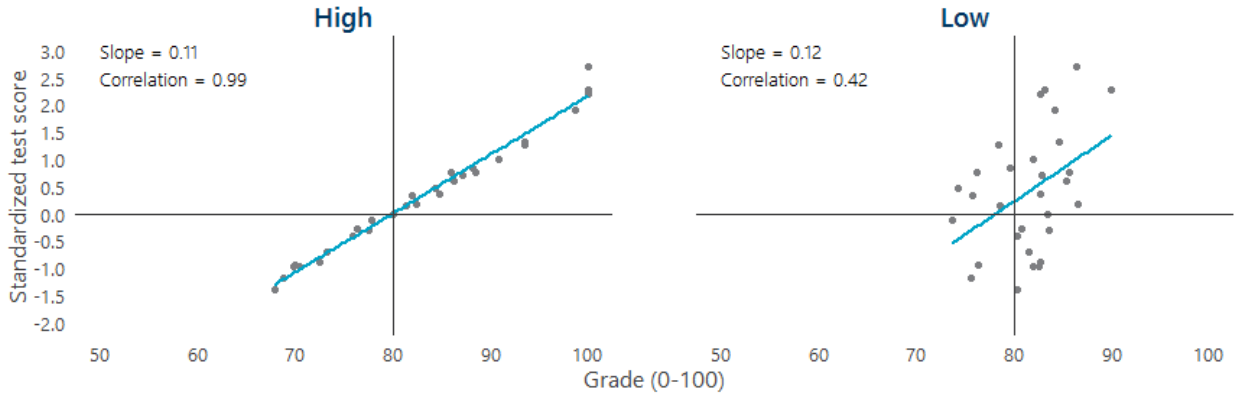
Tables and figures

Figure 1: Comparing different test-grade associations for identical set of 30 students' test scores.

High vs. Low SEVERITY



High vs. Low ALIGNMENT



High vs. Low GUARANTEE

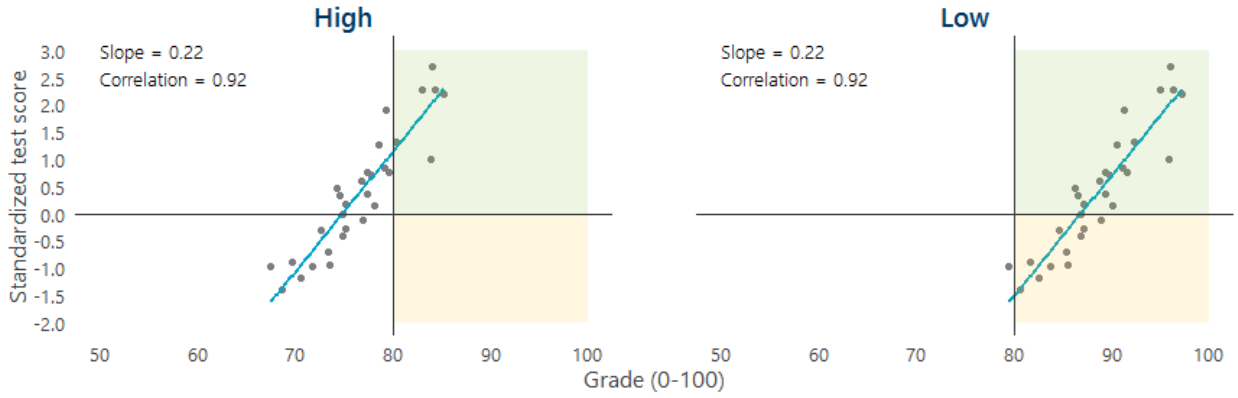


Table 1: Percent of 6th-8th grade students meeting grade-level test expectations by grade earned

	District A	District B	District C
ELA			
A	53%	77%	64%
B	20%	51%	22%
C	6%	12%	7%
D or Lower	3%	5%	3%
Math			
A	47%	67%	70%
B	16%	35%	22%
C	4%	5%	4%
D or Lower	1%	1%	1%