Estimating High School GPA Weighting Parameters with a Graded Response Model

John Hansen
Harvard Graduate School of Education
Cambridge, MA
john_hansen@mail.harvard.edu

Philip M. Sadler, Gerhard Sonnert
Harvard-Smithsonian Center for Astrophysics
Cambridge, MA 02138

July, 2017

PRELIMINARY DRAFT
Abstract

The high school grade point average (GPA) is often adjusted to account for nominal indicators of course rigor, such as “honors” or “Advanced Placement.” Adjusted GPAs—also known as weighted GPAs—are used widely for computing students’ rank in class and in the college admission process. Despite the high stakes attached to GPA, weighting policies vary considerably across states and schools, with little empirical justification. We discuss the suitability of the graded response model for estimating GPA weighting parameters on a latent scale, and demonstrate how to transform the latent-scale parameters to a conventional GPA scale to inform GPA weighting policy.

Key words: Graded response model; item response theory; higher education; admissions
High school grade point average (GPA) is a high-stakes measure of students’ academic achievement. In the college admission process, a small change in GPA can have significant consequences, especially in universities using a high school rank or GPA threshold for automatic admission or disqualification (Horn & Flores, 2003). Zimmerman (2014) found that students who barely qualified for admission to public universities in Florida earned more later in life than otherwise similar students who barely missed qualifying based on their GPAs. Shifts toward test-optional and class rank-based admission policies in recent years suggest that the importance of high school grades may be increasing (Hiss & Franks, 2014).

Atkinson and Geiser (2009) concluded that researchers generally agree that academic success in high school coursework is the best predictor of college success. This consensus among scholars is reflected among college admission professionals, who have ranked high school grades in college preparatory courses as the top factor in college admission decisions for decades (Clinedinst, Koranteng, & Nicola, 2015). Yet, there is no consensus about how several years of high school grades across a heterogeneous set of courses should be aggregated into a single number. In California, high school transcripts often include several GPAs, each computed using a different approach: For admission to the University of California or California State University system, advanced courses receive an extra grade point (UC-CSU, 2016); to determine eligibility for the Cal-Grant, the applicant’s first and last year of high school and all physical education courses are excluded, and no extra points are awarded for advanced courses (Cal-Grant, 2016); to assign class rank within a high school—which has implications for college admission in California—schools are free to choose their own approach. In North Carolina, until 2015, awarding two bonus grade points for Advanced Placement (AP) courses was a statewide policy (i.e., a B in an AP course received 5.0 grade points, while an A in a standard course received 4.0).
(North Carolina, 2014, 2015). In Miami-Dade Schools, AP course grades of A or B receive two bonus points, and a course grade of C receives one bonus point (Miami-Dade, 2016).

Two fundamental problems with using high school GPA as a high-stakes measure of student achievement are: (1) Not all students take the same courses, and (2) grades are ostensibly measured on an ordinal—not interval—scale. As shown by Arrow (1951) and discussed by Vickers (2000) in the context of GPAs, there is no optimal method for aggregating ordinal measures into a single composite. To compare academic achievement across students who took different courses from different teachers at different times, letter grades are typically transformed to numbers and averaged, yielding a GPA. In practice, Lang (2007) found that the most common aggregations were unweighted averages (all courses are weighted equally), weighted averages (e.g. where some courses receive a weight of zero), and rigor-adjusted averages that award extra points for nominal designations of course rigor (e.g. “Honors” or “Advanced Placement”). For a student who took $n$ high school courses, where $w_i$ denotes the scalar weight for course $i$’s contribution to the GPA, one can express the weighted GPA equation as:

$$ Weighted\ GPA = \sum_{i=1}^{n} w_i(GRADE_i) \quad (1) $$

The incorporation of a measure of course rigor by awarding bonus points for a subset of courses can be expressed as:

$$ Rigor\ Adjusted\ GPA = \sum_{i=1}^{n} w_i(GRADE_i + BONUS_i) \quad (2) $$

The term weighted GPA is commonly used to refer to rigor-adjusted GPAs, and there is no common designation for a GPA wherein some courses receive weights of zero. Many states and high schools include both a scalar-weight adjustment (e.g. for physical education courses, $w_i = 0$) and a bonus adjustment for advanced courses. For sake of simplicity, hereafter we use the term
weighted GPA to refer collectively to GPAs with course-varying scalar weights and/or bonus point adjustments.

In this study, we discuss and demonstrate the use of the graded response model (Samejima, 1969) for constructing or analyzing GPA weighting policies. We focus on three kinds of validity evidence: internal coherence, criterion prediction, and consequences for equity. The graded response model (GRM) is appealing because its underlying assumptions are compatible with conventional GPAs, it uses the same data as conventional GPAs, and GRM-estimated parameters are conceptually similar to conventional GPA weighting parameters. This article is organized into four sections. Section 1 briefly reviews previous research on the validity of the high school GPA. Section 2 discusses previous examples of GRM applications to course grade data and the theoretical suitability of using a graded response model to study high school GPA weighting parameters. Section 3 demonstrates how a GRM-based analysis could be used to inform policies on GPA adjustments for course rigor. Section 4 summarizes and concludes.

1. Previous Research on the Validity of the High School GPA

A great deal of research has been done on grading in schools. Brookhart and colleagues (2016) reviewed over 100 years of research about grading, and they concluded that the construct of interest measured by grades is multidimensional, consisting of “cognitive and non-cognitive factors reflecting what teachers value in student work” (p. 1). A key distinction between validity research on grades and standardized test scores is that multidimensionality is typically recognized as a feature of the former and a flaw in the latter. Sawyer (2013) and Atkinson and Geiser (2009) discussed high school grades and standardized testing in the context of postsecondary preparation and the college application process. Both studies emphasized that predictive validity is one of many factors that merit consideration. Others include diagnostic
utility, face validity, signaling to students (Atkinson & Geiser, 2009), and institutional objectives and mission (Sawyer, 2013). In this study, we address validity issues that arise through post-hoc grade adjustments and aggregations, not the initial grade-generating process.

Many studies of the high school grade point average have shown that high school GPA predicts first-year college grades (e.g. Kobrin et al., 2008), cumulative college GPA (e.g. Geiser & Santelices, 2007), and rates of college graduation (e.g. Bowen et al, 2009). The merits of GPA relative to standardized test scores is an ongoing subject of research, particularly vis-à-vis the influence of socioeconomic status (SES) (Atkinson and Geiser, 2009; Rothstein, 2004; Zwick, 2007; Zwick and Himelfarb, 2011)

Studies have also found that students who take advanced courses in high school, such as AP courses, tend to outperform students who do not (e.g Keng & Dodd, 2008). Geiser and Santelices (2004), Klopfenstein and Thomas (2005), Sadler and Tai (2007, 2010), and Warne (2017) found that the strength of the relationship between advanced course participation and college achievement depends on the additional covariates in the model. This line of research raises the question to what extent GPAs should be weighted by course type.

Whether unweighted or weighted GPAs better predict college performance—and how the relationship varies by postsecondary outcome or the inclusion of additional covariates—is not well-established. Warne et al. (2014) found that unweighted GPAs better predicted college GPAs for a sample of 700 pre-med students in Texas. Evidence of the consequences of grade weights is scant as well. The most relevant research is Klopfenstein and Lively’s (2015) study of changes in GPA-weighting policies in Texas. The authors found that when high schools increased the weights assigned to AP courses, relative increases in AP course participation were largest among students who did not qualify for free or reduced-price lunch. From a validity
standpoint, increased enrollment in AP courses as a result of a change in weighting policy is not necessarily problematic, even if the increases were largest for higher-SES students. The question is how well the post-change GPAs support valid inferences compared to the pre-policy change GPAs. If the policy change led students to take harder courses, learn more, and perform as well in college as one would expect given their GPAs, then one would conclude that the measure supported the same inferences before and after the policy change. (Though the question of whether the consequences of the re-weighting were desirable in terms of student tracking or college admission outcomes would remain.) A useful framework for analyzing the consequential validity of GPA weights is Koretz’s (2006) “score inflation,” which characterized test score inflation as an increase in high-stakes scores absent a concomitant increase in the construct of interest. Using Klopfenstein and Lively (2015) as an example, if the policy changes in Texas led to GPAs that systematically over-predicted college performance for all students, that would be evidence of inflation. Uniform inflation across all students may be undesirable, but whether it does any harm depends on the consequences of the over-prediction. If the consequence of interest in Texas is whether students are in the top 10% of their graduating class, and if each student’s GPA within a high school increases by 0.05 grade points, no harm is done because rank order is preserved. However, if the inflation is non-uniform—meaning that it is differentially larger for some groups than others—the potential consequence of the weights threaten the validity of the new weighted GPA measure. Any differential inflation within or between schools raises concerns about consequential validity.

*Criteria-Based Validation of Weights for Advanced Courses*

One approach to estimate GPA weights for advanced courses is to rely on an external criterion. Although they reach different conclusions, Geiser and Santelices (2004) and Sadler and
Tai (2007) employed similar, criterion-based empirical strategies. They used multiple regression where the outcome is a measure of college achievement, and the independent variables included indicators for participation in advanced coursework. Sadler and Tai (2007) based their recommendations of 0.50 bonus grade points for honors courses and 1.0 bonus grade points for AP courses on the relationship between high school science course participation and college science course performance. They compared the final grades in an introductory college science course among students who did and those who did not take honors or AP versions of the same course in high school. Using college chemistry as an example, they fit the following ordinary least squares (OLS) regression model:

\[
COL\_CHEM = \beta_0 + \beta_1(HS\_GRADE) + \beta_2(HON\_CHEM) + \beta_3(AP\_CHEM) + \epsilon,
\]

where \(HON\_CHEM\) and \(AP\_CHEM\) are dichotomous indicator variables, and standard chemistry is the comparison category. \(HS\_GRADE\) is the letter grade earned in a student’s last high school chemistry course. In some specifications they also added indicators for college instructor and included demographic control variables and SAT scores. The college course performance differentials uniquely explained by honors and AP course performance were translated to a grade scale by identifying the magnitude of a high school grade \((HS\_GRADE)\) increase equivalent to the differential. Using coefficients from a regression represented by Equation (3), the recommended bonus point adjustment for taking an honors chemistry course was:

\[
Bonus = \frac{\beta_{HON\_CHEM}}{\beta_{HS\_GRADE}}.
\]

To use the Sadler and Tai (2007) methodology—or a similar OLS regression approach—to inform a bonus joint adjustment policy, the first challenge is selecting a criterion. For example, one could use all college grades, first-year college grades, total credits accumulated, persistence
beyond the first year, or degree completion. In selecting college science course performance as a
criterion, Sadler and Tai (2007) implicitly assumed that college science course performance
suffices as a scale for measuring the equivalency between $\beta_{\text{HON, CHEM}}$ and $\beta_{\text{HS, GRADE}}$. In this
framework, the correct high school grade adjustment is the one that minimizes prediction error in
college course performance, either because one cares about predicting college course
performance for its own sake or because college course performance is considered a sufficient
measure of a broader construct of interest. Either way, the assumption is that high school grades
are meaningful insofar as they predict college grades in the courses students choose to take, at
the colleges where they are admitted and choose to enroll.

A second challenge for criterion selection is choosing what kind of variability in the criterion
is the “correct” variability for identifying grade weights. Because many factors other than
advanced course participation can explain variability in college performance, researchers have
tried to isolate variance attributable to advanced course participation by using covariates to
restrict the criterion variance of interest to within-college variance (Kloptenstein & Thomas,
2005), within-college-by-instructor variance (Sadler & Tai, 2007), and/or variance unexplained
by demographics (Rothstein, 2004). Rothstein (2004) discussed selection bias and modeling
strategies in detail, especially the issue of non-random selection of students into colleges and
courses. An equivalent way to frame this challenge is as one of model selection. Model selection
is consequential because many factors are correlated with high school grades, participation in
advanced courses, and college grades. Ultimately, different model specifications support
different inferences and different interpretations of the construct. The criterion selection problem
motivates the use of the GRM, since the GRM relies only on observed measures of high school
performance—the exact data one needs to compute a conventional GPA, and nothing more.
2. Graded Response Model GPAs

The first published application of a graded response model (GRM) to linking grades from different courses onto a common scale is Young’s (1990a) study of college performance among undergraduates. He found that the correlation between college GPA and SAT scores for entering students increased substantially when using a GRM-scaled college “GPA,” instead of the conventional GPA. Similar to Young (1990a), most studies of alternative GPA calculations have focused on correcting the college grade point average for differential grading stringency across college courses and departments (Caulkins et al., 1996; Stricker et al., 1994; Young, 1990b).

Conceptually, the most similar empirical analysis to this study is Bassiri and Schulz’s (2003), wherein the authors use ACT performance as an anchor to adjust the high school GPA to account for different grading practices between high schools.

The present study’s application of a GRM is distinct from previous work in that its primary focus is linking difficulty across grade-by-course combinations, rather than finding a scale to optimize prediction of a specific criterion. We do not explicitly seek the estimator of student skill that best predicts student success in college. We choose not to incorporate standardized achievement measures (e.g. SAT, ACT, or AP scores) in our model—although it would be straightforward to do so—as these additional tests plausibly measure a construct that differs from the one measured by the high school GPA (Willingham et al., 2002).

The appeal of estimating high school course rigor with a GRM is that it uses the exact same data as a conventional GPAs and no additional data, retains the most important assumptions of conventional GPAs, and relaxes non-essential assumptions. Similar to any conventional GPA, a GRM measure of a student’s high school achievement: (1) Treats the construct of interest underlying all of a student’s grades (hereafter $\theta$) as one-dimensional; (2) assumes students who score A’s in a given course tend to have higher levels of $\theta$ than students who score B’s in the
same course (and the same for B’s over C’s, etc.); (3) assumes that the construct of interest is identifiable from observed grades received in courses taken, and that course-taking patterns—including non-existent or “missing” grade data—can be ignored. In item response theory, this is the conditional independence assumption. Rubin (1976) classifies data that are missing at random conditional on observable data as missing at random (MAR). Assuming data are MAR implies that a student who has no grade in a certain course does not necessarily have a lower or higher \( \theta \) than one who has a grade in the course. The specific set of courses taken can be ignored without introducing bias into estimates of \( \theta \). Setting aside their potential flaws, we treat these assumptions as fundamental assumptions of GPAs. In short, the trait of interest is unidimensional, higher grades imply higher trait values, and the trait is identifiable from observed data only. A potential benefit of estimating a “GPA” with a GRM is that the GRM relaxes three additional constraints of conventional GPAs. In the case of the GRM, (4) Differences in course difficulty are estimated, not presupposed; (5) the trait’s scale is estimated from the data without assuming an interval property (e.g. the difference on the scale between an A and a B is not assumed to be the same as the difference between a B and a C, and these distances are not assumed to be constant across courses); and (6) all courses are not assumed to be equally reliable measures of \( \theta \). Theoretically, relaxing these constraints will improve the internal coherence of the measure unless the imposed constraints indeed conform to the data generating process.

To be clear, the GRM estimates parameters with conventional GPA weighting analogs, but it does not directly estimate them on the conventional GPA scale of Equation (1) or (2). The GRM assumes the existence of latent variable, conventionally assumed to be normally distributed with a mean of zero and unit variance. In estimating parameters, the model finds the \( \theta \) scale using a
cumulative log-odds principle (Ostini & Nering, 2005). Formally, the equation for estimating the
GRM can be expressed as:

\[ P(Y_{ij} \geq k | a_i, b_{ik}, \theta_j) = \frac{1}{1 + \exp \{-a_i(\theta_j - b_{ik})\}} \]  

(5)

where \( Y_{ij} \) is the letter grade earned by student \( j \) in course \( i \); \( k \) represents the possible letter grades;
\( a_i \) can be interpreted as the precision with which course \( i \) measures \( \theta \); \( \theta \) is a latent variable one
could interpret as student academic skill (or the construct one supposes is measured by grades in
high school courses), and \( b_{ik} \) is a measure of the difficulty of achieving letter grade \( k \) in course \( i \).
For each course, \( k - 1 \) threshold parameters are estimated. Each \( b_{ik} \) is the \( \theta \) value in which a
student for whom \( \theta_j = b_{ik} \) has a .50 fitted probability of earning letter grade \( k \) or higher in course
\( i \). Comparing estimates of \( b_{ik} \) allows one to link difficulties of letter grades across different
courses. If \( Y_{ij} \) is truly a function of \( \theta \), item parameters, and nothing else, then, for example,
identical \( b_{ik} \) parameters for the probability of earning a B- or better in AP Statistics and earning
an A- or better in precalculus can be interpreted as evidence of equal difficulty of earning at least
these grades in the respective courses. A stronger case for equivalence in difficulty for letter
grades across two courses would require the courses to have identical values of \( a_i \). In this case,
fitted probabilities of exceeding the thresholds would be equivalent for all values of \( \theta \), not only
where \( \theta_j = b_{ik} \). More precisely, \( a_i \) parameterizes the rate at which for course \( i \) the probability of
exceeding thresholds changes as a function of \( \theta \). If the probability of exceeding thresholds
changes little as a function of \( \theta \) for course \( i \), the lower \( a_i \) indicates that student grades in course \( i \)
are less informative measures of \( \theta \).

**GRM Parameter Interpretation in the Weighted GPA Context**
Figure 1 is a stylized representation of perfect correspondence between a common weighted GPA policy and grade-by-course difficulty parameters from a graded response model. The most common policy for adjusting grades in honors and AP courses is to award an additional letter grade for AP courses and one half of a letter grade for honors courses (Lang, 2007), which we plot in Panel A. Panel B of Figure 1 illustrates how grade-by-course difficulty parameters can be converted between a weighted GPA scale and the GRM’s latent variable scale. For sake of illustration, we assume a θ scale where the average student earns a B or higher in 50% of standard courses. In this scenario, where the GRM estimates conform perfectly to the weighted GPA policy, the GRM boundary parameters \( b_k \) for common mathematics and science courses are a linear shift of \(-3.0\) units from the weighted GPA scale. Under these stylized conditions, note that one can transform students’ θ to a 5.0 GPA scale simply by adding three points and truncating GPAs above 5.0 or below zero.
Figure 1. Conceptual model for GRM-estimation of high school GPA weighting parameters at the letter-grade-by-course level. (A) Weighted GPA policy in which honors courses receive 0.50 bonus grade points and Advanced Placement courses receive 1.0 bonus grade points. (B) Probability of earning a given course grade or higher as a function of $\theta$.

Panel B illustrates the relationship between boundary parameters ($b_{ik}$) and grade-by-course difficulty. In this example, the $b$ parameters for a B in AP Statistics and an A in standard statistics are the same. Because $b_{B,AP\text{-STATS}} = b_{A,STATS}$, the curves intersect where $p=.50$. For student $j$ such that $\theta_j = b_{B,AP\text{-STATS}} = b_{A,STATS}$, the probability of earning a grade of B or higher in AP Statistics is the same as the probability of earning an A in a standard statistics course ($p=.50$). Hence, these grade-by-course combinations are equally difficult to achieve for student $j$.

Panel B also illustrates that the $a_i$ parameters are different for standard Statistics and AP Statistics, which means that whether it is more or less difficult to earn a B in AP Statistics or an
A in standard Statistics depends on $q$. Nevertheless, the differential is symmetric around $b$, implying equivalent difficulty on average (or, more precisely, equivalent average difficulty for any symmetric distribution of students centered at $b$). Biology illustrates a case where the $a_i$ parameters are the same for AP Biology and standard biology but the $b$ parameters differ by 0.33 units of $\theta$. As a result, for any fixed value of $\theta$, the probability of earning a D or higher in AP Biology is greater than the probability of earning a C+ in standard biology.

In this hypothetical example, a difference of 0.33 units on the $\theta$ scale is equivalent to 0.33 grade points on a 5-point GPA scale. In general, the magnitude of similar differences on the $\theta$-scale will be sensitive to rescaling, even though the scale itself is arbitrary up to a linear transformation. Following Lord (1980), consider a transformation of the current $\theta$ scale to an alternative scale with greater variance, such that the new scale has $m$ times the standard deviation of the old ($m\sigma_{\text{OLD}} = \sigma_{\text{NEW}}$). Fitted probabilities are preserved so long as $a_i$ and $b_{ik}$ are rescaled as well.

$$P(Y_{ij} \geq k|a_i b_{ik}, \theta_j) = \frac{1}{1 + \exp \{-a_i (\theta_j - b_{ik})\}} = \frac{1}{1 + \exp \{-\frac{a_i}{m} (m(\theta_j - b_{ik}))\}}$$

(6)

Graphically, this is equivalent to multiplying by $m$ the $x$ axis labels in Figure 2. Yet, despite the substantive equivalence of the models, the magnitude of the difference on the $\theta$-scale is not preserved: $(mb_D^{AP,Bio} - mb_C^{Bio}) \neq (b_D^{AP,Bio} - b_C^{Bio})$. We revisit this issue in Section 3.2.

As always, substantive interpretations of parameter estimates rely on the assumption that the GRM model fits the data. As Thissen (2016) argues, the binary answer to the “bad question” of whether an item response model fits is that is does not. Certainly, in this case, there are many reasons why $Y_{ij}$ might not be exclusively a function of item parameters and $\theta$. A GRM does not ameliorate all shortcomings of grades or conventional GPA aggregation approaches. Many
salient criticisms of the GRM approach apply equally to conventional GPAs. Unless explicitly modeled, neither accounts for variability in grading standards in nominally identical courses offered by different teachers, in different high schools, or on different occasions. The extent to which these omissions are flaws or features depends on the relevant inference and how one defines the construct that GPA is supposed to measure. Concerns that apply similarly to GRM and conventional GPAs—and possible ameliorations—merit additional study, but we do not address them comprehensively here. Our claim is that a GRM approach is useful and well-suited to estimating how grades received in different courses can be compared to one another on a common scale. The parameters of interest for course rigor analysis are $b_{ik}$, and their interpretation is analogous to canonical GRM applications under the following conditions: (1) Grade by course difficulty parameters measure grading stringency net of course-specific differences in student effort, and (2) $\theta$ is treated as time-dependent.

Interpreting $b_{ik}$ parameters as measures of grading stringency net of course-specific student effort is consistent with traditional GRM parameter interpretations. Still, it is worth making the distinction that course rigor is operationalized as a property of an observed grade distribution, not an unobserved effort distribution. Unobserved heterogeneity in effort across items of different difficulty levels is not a problem in GRM applications so long as one recognizes that identical $b_{ik}$ parameters do not imply equality of effort. The GRM estimates the probability that students with a given $\theta$ would receive a grade of $k$ or higher if they were to take the course. It allows one to identify dissimilar patterns of $Y_{ij}$ that imply the same $\theta$, not whether students who earned a B in AP Statistics could have earned a B+ in pre-calculus with equal effort. If grading stringency as manifest in grade distributions—conditional on $\theta$—does not vary by course, then neither will $b_{ik}$ parameters.
Unlike most educational measures, the high school GPA is an aggregation of measures from a several year period. Time is not explicitly modeled, and $Y_{ij}$ is theoretically a function of $\theta_i$, item parameters, and *when students took the course*. A student who excelled in calculus at 12$^{th}$ grade would not necessarily have fared well in ninth grade, before they had studied advanced algebra and trigonometry. One resolution is to reinterpret $\theta$ as a time-adjusted measure of academic skill: the propensity of a student to succeed in coursework for which they have the necessary time-dependent academic skill to take. As a result, ninth graders’ $\theta$ would estimate how well they would do in calculus when prepared to take it, but it would not measure how well they would perform in calculus were they to take it in lieu of geometry in ninth grade. In the context of the high school GPA and nominal course rigor adjustment, this approach does not penalize students whose first high school math class is algebra 1 (likely putting AP calculus by 12$^{th}$ grade out of reach). It would presume that a student who took precalculus in 11$^{th}$ grade and chose not to take calculus the following year can be compared in terms of $\theta$ to a student who chose to take AP calculus in 12$^{th}$ grade. Ultimately, while this interpretation departs from canonical GRM applications, it is entirely consistent with conventional interpretations of the GPA construct.

Overall, using a GRM to estimate letter grade-by-course difficulty parameters is similar to a linking study using non-equivalent groups, a common-item design, and concurrent calibration (Kolen & Brennan, 2004). The estimation of model parameters requires the existence of multiple courses and crossing of students and courses. Common items—known as anchors—are used to link disparate groups of students to one another on a common scale. The quality of the link depends heavily on the anchors. If the only crossing of high-ability and low-ability students occurred in Algebra 2, and all higher-ability students were assigned to a teacher who graded
more stringently, one would not be able to distinguish the high-ability students and the rigor of their other courses from the low-ability students and the rigor of their courses.

3. Demonstration

In this section, we fit a GRM model to actual course grade data and illustrate how a study of the validity of course rigor weights could proceed. To be clear, we do not present this analysis as conclusive evidence of any particular weighting scheme’s universal characteristics. The results of similar analyses may differ considerably across samples. Theoretically, if the model fits the data, parameters should be sample invariant up to a linear transformation, but the extent to which this holds in practice is an empirical question beyond the reach of a one-sample study. In addition to comparing our estimated grade-by-course difficulty parameters to policy-dictated parameters for computing GPAs, we compare the GRM-weighted GPAs to other simulated weighted GPAs in terms of college grade prediction and bias favoring socioeconomically advantaged students. Estimates of bias are comparative because a student’s true GPA is not known. We use parental educational attainment as a proxy for socioeconomic status (SES), because it is the only one of the “Big 3” measures of SES available in the data (NCES, 2012). Predictive power and SES-attributable bias are perhaps the two most scrutinized features of measures used for college admission, and we characterize the former as correlational validity evidence and the latter as consequential validity evidence.

3.1 Data and the GRM Model

The data for this section of the study come from the Factors Influencing College Success in Mathematics survey (FICSMath), which is funded by the National Science Foundation. A stratified national random sample of 10,492 first-year college calculus students was given the survey at the beginning of the course. Most of the survey items identify pedagogical decisions
made by teachers that may contribute to later student success in mathematics and science. It also includes items about a student’s high school, demographic background, and, most importantly for this study, the mathematics courses they took in high school and the letter grades they received. At the end of the term, the calculus instructor reports each student’s grade from the course. For the analytical sample, in this order, we listwise delete students for whom no final grade was reported (359), students who reported taking fewer than three quantitative courses (1535), and students who did not respond to the survey item on parental educational attainment (262). We focus only on quantitative courses (mathematics courses plus AP Physics), because the survey has very limited coursework data on non-quantitative coursework and a GRM assumes a unidimensional construct. The analytical sample contains 8,336 students. Table 1 shows that students taking calculus as first-year college students are a relatively high-performing group of students. The average unweighted GPA for quantitative courses in the analytical sample is 3.50.

Table 1

<table>
<thead>
<tr>
<th>Descriptive Statistics for the Analytical Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Analytical Sample (N = 8336)</strong></td>
</tr>
<tr>
<td>Calculus Grade</td>
</tr>
<tr>
<td>GPA: No Weights</td>
</tr>
<tr>
<td>GPA: GRM Weights</td>
</tr>
<tr>
<td>Parent has Bachelor’s</td>
</tr>
<tr>
<td>Total Course Grades Reported</td>
</tr>
<tr>
<td>Honors Course Grades Reported</td>
</tr>
<tr>
<td>AP Course Grades Reported</td>
</tr>
</tbody>
</table>

Note. All data above except Calculus Grade were self-reported by students participating in the Factors Influencing College Success in Mathematics study (FICSMath). First-year college calculus classes were sampled from a randomly selected set of U.S. colleges during the 2009-2010 academic year.

Figure 2 plots the GRM-estimated grade-by-course difficulty estimates on the θ scale. For this sample, which is restricted to students enrolling in first-year college calculus, the ranking of GRM course difficulty parameter estimates \( b_{ik} \) generally supports the practice of GPA weights.
for AP courses. The proportion of A’s received was lowest in AP Calculus BC and second lowest in AP Calculus AB (Table A1). Students who report taking AP Calculus tend to have higher \( q \) estimates (\( \Delta = 0.34, t(1) = 15.82, p < .001 \)), leave-out-calculus GPAs (\( \Delta = 0.19, t(1) = 14.90, p < .001 \)), and SAT scores (\( \Delta = 53.3, t(1)=21.02, p < .001 \)) than students who do not. The lower rate at which students who have high grades in other classes receive A’s in AP Calculus provides the GRM’s empirical basis for identifying the course as particularly difficult.

\[ \text{Figure 2. Each point indicates the mathematics skill (} \theta \text{) necessary to have a 0.50 probability of earning the given letter grade or higher in the given course. Confidence intervals are } \pm 1 \text{ SE. Grades of D- and D are collapsed into a single category due to sparse data at the low end of the grade distribution.} \]

Overall, the results presented in Figure 2 generally support the practice of GPA adjustments of some magnitude for nominal indicators of course difficulty. However, in contrast to weighting schemes in which AP courses receive one full letter grade adjustment, earning a B+ or better in
the most difficult AP course appears similarly difficult as earning an A in a standard algebra
course (for a student with a \( \theta \) around -0.25). In most cases, one third of a letter grade is closer to
the course letter grade adjustment to create equivalent letter grade difficulties between AP and
non-AP courses. For honors courses, the differential is typically less than one third of a letter
grade. Parameter estimates for honors and standard versions of courses in the same subject are
very similar. Even if the above \( \theta \)-scale were linearly transformed to a conventional 4.0- or 5.0-
point GPA scale, equivalent \( b_{ik} \) parameter estimates would remain equivalent. An unexpected
feature or Figure 2 is the large “step” between B and B+ grades, and between C and C+ grades,
which is present in every course. Consequently, achieving a B in AP Calculus does not appear to
be more difficult than earning a B+ in most other mathematics classes. This is inconsistent with
simple bonus point adjustment policies characterized by Equation (2), since it suggests that high
school letter grades are awarded on a scale that lacks the interval property.

3.2 Adapting Latent-scale Parameters to a Familiar GPA Scale for Use in Policy

Even for a state or district with high quantitative research capacity, the benefits of wholesale
transition to GRM-estimated GPAs seem unlikely to justify the practical costs and confusion. To
be clear, we do not suppose that full-fledged adoption of GRM-GPAs is advisable. Our supposed
use of the GRM is estimating \( BONUS_i \) parameters for use in a policy characterized by Equation
(2). While \( BONUS_i \) parameters from Equation (2) are not directly estimated by the GRM, the \( \theta \)-
scale parameter estimates can be rescaled to accord with a conventional GPA scale.

Consider a policy context wherein \( BONUS_i \) is fixed at zero for standard courses and
potentially non-zero for honors and AP courses, and the objective is to estimate bonus point
adjustments for honors and AP courses on a weighted GPA scale. Also assume that the standard
deviation of the target GPA scale is not known. In this case, a possible mean-sigma (Kolen &
Brennan, 2004) approach would be to identify the linear transformation that maps the $b_{ik\theta}$ for standard courses to a known GPA scale (i.e. the conventional 4.0 scale for standard courses), apply the same transformation for honors and advanced courses, and set policy for $BONUS_{HON}$ and $BONUS_{AP}$ using GPA scale differences in $b_{ik}^{GPA}$.

Specifically, let $m$ denote the sigma rescaling parameter and $z$ denote the mean re-centering parameter. In the typical case of mean-sigma linking, the means and standard deviations of two sets of parameter estimates are known. In this case, we have one set of parameter estimates for all courses on the $\theta$ scale, $b_{ik\theta}$, and one set of known, policy-dictated parameters for letter grade point values in standard courses, $b_{k}^{GPA\ast}$ (the $b_{k}^{GPA\ast}$ vector contains the standard course grade point values plotted in Panel A of Figure 1). This allows one to use an alternative estimator of $m$ that exploits the known, fixed distance between letter grade values in standard courses on the target scale.

$$m = \frac{b_{k2}^{GPA\ast} - b_{k1}^{GPA\ast}}{\frac{1}{n} \sum_{l=1}^{n} b_{lk2}^{\theta} - b_{lk1}^{\theta}},$$  \hspace{1cm} (7)

where $b_{k2}^{GPA\ast}$ and $b_{k1}^{GPA\ast}$ are two fixed points on the known GPA scale, and $b_{lk2}^{\theta}$ and $b_{lk1}^{\theta}$ are the corresponding points on the $\theta$ scale for standard course $i$. For example, since $b_{ik}$ parameters are estimated imprecisely at the lower end of the distribution in our sample, we could estimate $m$ using the distance between C and A grades. In this case, the numerator of Equation (7) would be $b_{A}^{GPA\ast} - b_{C}^{GPA\ast}$, or 2.0, since 4.0-2.0 = 2.0. For the eight standard courses in our sample, the average difference between $\hat{b}_{iA}^{\theta}$ and $\hat{b}_{iC}^{\theta}$ is 2.55. Therefore, 2.0/2.55 is our estimate of $m$, the factor by which the $\theta$ scale is compressed (or stretched) to match the conventional GPA scale.

For our sample, Figure 3 shows the compression of points in Panel B compared to Panel A.
We estimate $z$ with the mean difference in $b_{ik}$ across the two scales for standard courses, after adjusting $b_{ik}^\theta$ by a factor of $m$. Continuing to use $b_{ik}$ for C through A as our scale anchors, the equation is:

$$z = \frac{1}{nq} \sum_{i \in I} \sum_{k \in K} b_{ik}^{GPA} - mb_{ik}^\theta,$$

where $I=$\{Algebra 1, Algebra 2, ..., Statistics\}, $K=$\{C, C+, ..., A\}, and there are $n$ elements in $I$ ($n = 8$ in our sample) and $q$ elements in $K$.

The mean-sigma equation for linking $b_{ik}^\theta$ to the GPA scale is $b_{ik}^{GPA} = m b_{ik}^\theta + z$. Once all course-by-grade difficulties are on the same scale, one can estimate BONUS quantities of interest. Whether the correct BONUS quantity of interest is the difference in difficulty between an AP course and standard course in the same subject ($\bar{b}_{AP,stats}^{GPA} - \bar{b}_{stats}^{GPA}$), or the difference in difficulty between an AP course and the average standard course ($\bar{b}_{AP,stats}^{GPA} - \bar{b}_i^{GPA}$)—or perhaps something else—is ultimately a policy question. Unweighted or weighted average differences in $\bar{b}_{ik}^{GPA}$ and the comparison course(s) are potential BONUS estimators. Possible candidates for weights could be the number of students taking the course or a function of the standard error for $\bar{b}_{ik}^{GPA}$.

In our sample, for demonstration, we estimate the unweighted average difference in $b_{ik}^{GPA}$ between the average AP course and the average standard course, continuing to restrict $k$ to letter grades C through A. In this case, the equation is:

$$BONUS_{AP} = \frac{1}{cq} \sum_{p \in P} \sum_{k \in K} b_{pk}^{GPA} - \frac{1}{nq} \sum_{i \in I} \sum_{k \in K} b_{ik}^{GPA},$$

(9)
where $P=\{AP \ Physics, \ AP \ Statistics, \ AP \ Calculus \ AB, \ AP \ Calculus \ BC\}$, and there are $c$ elements in $P$ ($c = 4$ in our sample). Our estimate of $BONUS_{AP}$ is 0.25 letter grade points, and the analogous estimate for $BONUS_{HON}$ is 0.02 letter grade points.

Figure 3 provides a visual illustration of the scale transformation for calculus and statistics only. Panel A plots $\beta_{ik}^\theta$, and Panel B plots $\beta_{ik}^{GPA}$. If the $\beta_{ik}$ for AP and standard courses were the same, they would fall on the main diagonal in both panels. The “1.0 Point Bonus Line” indicates where the $\beta_{ik}$ would fall if we found that Advanced Placement courses in statistics and calculus were 1.0 GPA points more difficult than standard courses in the same subject. If the $BONUS$ quantity of interest were the difference in difficulty between an AP course and standard course in the same subject, a simple estimator for $BONUS_{AP \ STATS}$ would be the mean vertical distance between the letter grades labeled “stats” and the main diagonal in Panel B.

![Figure 3](image_url)

*Figure 3.* Comparison across scales of estimated $b_{ik}$ parameters in calculus and statistics for standard and Advanced Placement courses. For calculus, AP Calculus AB is plotted. (A) $\theta$ scale. (B) GPA scale.

### 3.3 Correlational and Consequential Validity

To demonstrate how a GRM could support an analysis of the validity of GPA weighting policies, we investigate the degree to which awarding bonus points for advanced courses
systematically increases GPAs for higher-SES students beyond empirical justification in academic skill. We proceed in two parts. First, we estimate how well weighted and unweighted high schools GPAs explain college performance in a first-year calculus class. We address the role of SES in prediction by comparing the fit of regression models that do and do not account for the role of SES. This approach allows us to observe the extent to which any improved prediction of course performance due to weights is a function of SES (because higher-SES students take more advanced courses) or a better measure of student academic skill, after accounting for SES. Second, we estimate the extent to which differences in weighted GPA by SES are attributable to student academic skill. We interpret higher GPAs for higher-SES students due to GPA weights, net of academic skill, as evidence of differential GPA inflation (i.e. bias).

In both parts we compare the same sets of high school GPAs:

- **No Weights.** An unweighted GPA on a 4.0 scale.
- **GRM Weights.** Honors courses receive 0.5 bonus points and AP courses receive 1.0 bonus points, such that the maximum possible points in a course is 5.0. Weights in this range are fairly common and currently in use in Massachusetts (MA, 2013).
- **MA Weights.** Honors courses receive 1.0 bonus points and AP courses receive 2.0 bonus points, such that the maximum possible points in a course is 6.0. These are the largest weights of which we are aware, and they were used in North Carolina through the 2014-2015 school year (NC, 2015).

Table 2 shows high correlations between unweighted and GRM-weighted GPAs, and high correlations between MA-weighted and NC-weighted GPAs (~.95 Pearson correlation and ~.95 for the other weights).
Spearman Rank correlation). Correlations between these two pairs of variables are somewhat lower, particularly in the case of NC-Weighted GPA. Consistent with previous research, parental educational attainment is positively associated with honors and AP course participation, as reflected in the higher correlations between parental educational attainment and GPAs computed with bonus point policies.

Table 2

<table>
<thead>
<tr>
<th></th>
<th>Original Scale (Pearson)</th>
<th>Ranks (Spearman)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>None</td>
<td>GRM</td>
</tr>
<tr>
<td>No Weights</td>
<td>.94</td>
<td></td>
</tr>
<tr>
<td>GRM Weights</td>
<td>.91</td>
<td>.88</td>
</tr>
<tr>
<td>MA Weights</td>
<td>.77</td>
<td>.76</td>
</tr>
<tr>
<td>NC Weights</td>
<td>.05</td>
<td>.04</td>
</tr>
<tr>
<td>Parent Educ</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. ParBA is a dichotomous indicator for whether a student’s most educated parent graduated from college. ParEd uses an ordinal scale, which ranges from less than a high school degree to a graduate-level degree.

3.3.1. Predicting college calculus grades. The ability to predict a criterion is one form of correlational validity evidence. A problem with validating weights through their ability to improve the prediction of college grades, though, is that SES may also have a relationship with college grades, independent of academic skill. In this case, prediction models that omit SES would tend to over-predict grades for low-SES students. As a result, one could “improve” these prediction models by artificially lowering the academic measures of lower-SES students in order to “correct” the over-prediction these students’ performance. This is an absurd “correction,” of course. Relying entirely on college grades to validate a weighing approach risks the same error at the upper end of the SES distribution. If unweighted GPA systematically under-predicts college calculus grades for higher-SES students (and it does, as shown in Table 3), then artificially adding grade points to the GPAs of students with advanced courses (i.e. students who disproportionately come from higher SES families) will tend to partly correct for this systematic
under-prediction, regardless of how this underprediction comes about. Comparing prediction models with and without a measure of SES as a covariate—to block the direct path from SES to college grades—can illuminate the extent to which a given set of GPA weights results in a better model fit due to coincidence with SES, or independently of SES. Students potentially overperform or underperform expectations for many reasons, including factors related to SES, so a fair or SES-neutral weighting scheme is not necessarily the one that minimizes prediction error or minimizes the incremental predictive power of SES (which may also be exerted through other channels).

Our approach is comparative, and a reasonable comparison requires a defensible baseline. One can compare models with the weights in use in North Carolina and Massachusetts to a model with no weights, but a model with no weights is arguably an arbitrary reference point. It implicitly assumes that the correct weighting system weights all courses equally. We include a GRM-weighted GPA for precisely this reason. After computing—or one could say simulating—the weighted GPAs for all students, we fit Ordinary Least Squares (OLS) regression models of the forms:

\[
CALC\_GRADE = \alpha + \beta (GPA) + \epsilon \tag{10}
\]

\[
CALC\_GRADE = \nu + \gamma (GPA^\ast) + \nu \tag{11}
\]

\[
CALC\_GRADE = \zeta + \gamma (GPA) + \sum_k \lambda_k PARENT\_ED + \eta, \tag{12}
\]

where each model uses a GPA calculated according to a different set of weights. \(CALC\_GRADE\) is the grade earned by a student in college calculus. \(PARENT\_ED\) is a vector of dichotomous variables indicating the highest level of education attained by students’ parents, where the lowest level is less than a high school degree, and the highest level is a graduate degree. Note that Equation (11) uses \(GPA^\ast\), which is the variability in unweighted GPA that cannot be explained
by parental education attainment. Specifically, $GPA*$ is the vector of residuals ($\beta$) from a regression of the following form:

$$GPA_{NO\_WEIGHTS} = \tau + \sum_k \kappa_k PARENT\_ED + \mu$$  \hspace{1cm} (13)

The coefficient for GPA, $\gamma$, is the same by construction in Equations (11) and (12), but the error terms from the regressions are not the same. The quantities of interest from Equations (10-12) are the error terms, $\epsilon$, $\nu$, and $\eta$. The distribution of the error term is informative about the fit of the model, including its differential fit for different kinds of students. These are estimates of the prediction error one would expect in the population for colleges using the corresponding GPAs to predict first-semester calculus grades in analogous regressions. Comparing $R^2$ with and without accounting for parental educational attainment shows the extent to which incremental improvement in prediction attributable to GPA weights coincides with, or is independent of, the contribution of SES.

Table 3 shows that the GRM-weighted GPA has the strongest correlation with college calculus grade. The bivariate correlations range from .329 to .355 (the square root of the $R^2$ from Column 1), and the correlations for the unweighted and NC-weighted GPAs are the lowest. The specification in Column 2 excludes any influence of parental educational attainment. For the unweighted and GRM-weighted GPAs, $R^2$ decreases .002-.003 compared to Column 1. For the weighted GPAs, $R^2$ decreases by 0.004 and 0.005 for the MA-weighted and NC-weighted GPAs, respectively. Column (3) shows that the covariance of parental education and college calculus contributes approximately .008 to $R^2$ for all GPAs. This means that for the MA-weighted and NC-weighted GPAs, a larger proportion of their predictive power operates through covariance with parental educational attainment compared the unweighted and GRM-weighted GPAs. After partialling out SES, the NC-weighted GPA has the weakest correlation with college calculus
ESTIMATING GPA WEIGHTING PARAMETERS

grade. This illustrates the limitations of naïve bivariate correlation coefficients as prima facie evidence of validity. The GRM-weighted GPA has the highest correlation with college calculus grade despite its relatively weak with parental education compared to the other weighted GPAs. In practice, the difference between an $R^2$ of .120 compared with an $R^2$ of .125 may be insubstantial, but similarity along this dimension does not imply an absence of meaningful differences along other dimensions, as we show in the next section.

Table 3

Regression Estimates of the Association of College Calculus Grade with Weighted and Unweighted GPAs, With and Without Accounting for the Influence of Parental Educational Attainment

<table>
<thead>
<tr>
<th></th>
<th>Bivariate</th>
<th>GPA*</th>
<th>Parent Ed Controls</th>
<th>Bivariate</th>
<th>GPA*</th>
<th>Parent Ed Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(A)</td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
</tr>
<tr>
<td>GPA</td>
<td>9.40</td>
<td>9.313</td>
<td>9.313</td>
<td>5.774</td>
<td>5.725</td>
<td>5.725</td>
</tr>
<tr>
<td></td>
<td>(0.443)</td>
<td>(0.427)</td>
<td>(0.432)</td>
<td>(0.225)</td>
<td>(0.222)</td>
<td>(0.229)</td>
</tr>
<tr>
<td>$R^2$</td>
<td>.106</td>
<td>.104</td>
<td>.112</td>
<td>.124</td>
<td>.121</td>
<td>.130</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPA</td>
<td>8.566</td>
<td>8.442</td>
<td>8.442</td>
<td>6.379</td>
<td>6.266</td>
<td>6.266</td>
</tr>
<tr>
<td>(B)</td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
</tr>
<tr>
<td>GPA</td>
<td>(0.442)</td>
<td>(0.419)</td>
<td>(0.420)</td>
<td>(0.335)</td>
<td>(0.335)</td>
<td>(0.333)</td>
</tr>
<tr>
<td>$R^2$</td>
<td>.120</td>
<td>.116</td>
<td>.124</td>
<td>.107</td>
<td>.102</td>
<td>.110</td>
</tr>
</tbody>
</table>

Notes. The dependent variable in each regression is a student’s score in college calculus on a 100-point scale. GPA* is a measure of GPA constructed to be orthogonal to parental educational attainment. Coefficients for the set of parental educational attainment indicators from Model (3) are omitted from the table. Heteroskedasticity-robust standard errors, clustered at the college level, in parentheses.

3.3.2. **GPA inflation.** The question of interest in this section is: Are high school GPAs higher or lower than one would predict for high-SES students conditional on what we can observe about their academic skill, and to what extent do various GPA weights affect the answer? To address this, we fit a taxonomy of OLS regression models of the form:

$$GPA\_NO\_WEIGHTS = \alpha + \gamma(PARENT\_ED) + X\beta + \epsilon,$$  (11)
where the dependent variable is a measure of GPA and \(X\beta\) is a matrix of measures of academic skill and the corresponding regression coefficients. The necessary assumption that one must make to interpret \(\gamma\) as a measure of SES-related GPA inflation attributable to parental education is that \(X\beta\) sufficiently captures academic skill, and captures it equally well across the SES distribution. One can think of \(X\beta\) as potential empirical justification for why students with more educated parents have higher GPAs on average. For example, students having higher GPAs because they do higher quality work in school or have higher quantitative aptitude are potential justifications. Ideally one would have a “true” GPA on the right side of the equation, which would make additional variables unnecessary. We consider the GRM-GPA a reasonable estimate of the academic skill that GPA is supposed to measure due to its theoretical advantages and the empirical results in 3.3.1, showing that the GRM-weighted GPA was the best predictor of college calculus grades in all specifications despite its weaker correlation with SES, compared with the weighted GPAs. One could argue that the correct estimate of grade inflation for a given set of weights attributable to SES needs nothing further. A counter-argument is that the GRM estimates of academic skill systematically overestimate the skill of students in high schools with more lax grading standards, which may also have fewer honors and AP courses, which could lead to over-estimation of grade inflation. In this scenario, measures of student skill that occur outside the high school environment are worthy candidates for inclusion in order to account for differential grading standards across high schools. To address this concern, we fit models including SAT scores (or ACT scores converted to an SAT scale) and college calculus performance as well. These models will potentially underestimate inflation if parental education also systematically inflates SAT scores or college calculus grades. That is, if an SES measure systematically inflates a measure we use to test for SES-related inflation, we will underestimate
the magnitude of said inflation. This is perhaps the fundamental challenge of validating a measure; finding a flaw in one empirical measure requires the existence of a second measure without the flaw. Our estimates of GPA inflation for higher-SES students use increasingly broad interpretations of GPA, which plausibly underestimate inflation if one were to interpret more narrowly what a GPA should measure.

Table 4 shows that the simulated high school GPAs of students whose parents graduated from college are higher than measures of their academic skill suggest. Column (1) shows that, unconditional on academic skill, students with more highly educated parents have higher GPAs. In Columns (2-4), where $\theta$ is included as a covariate, the coefficients on the indicator for “Parent has BA” provide estimates of inflation on the GPA scale of the regression’s dependent variable. The specification in Column (2) estimates that unweighted GPAs are inflated for students with college-educated parents by 0.01 grade points on a 4.0 scale. Columns (3) and (4) add college calculus grade and SAT-Math score, which one may or may not consider relevant measures of the construct that the GPA should be measuring. We consider Columns (3) and (4) conservative estimates of inflation. Practically, after including only one or all available measures of academic skill—defining academic skill narrowly in relation to high school coursework performance or more broadly—the unweighted GPA has little SES-attributable inflation or deflation.

Panels B and C of Table 4 present coefficients for MA-weighted and NC-weighted GPAs. In these cases, the estimated GPA inflation attributable to weights for students with college-educated parents range from 0.04 points using Massachusetts weights with the most conservative specification, to 0.11 points for NC-weighted GPAs conditioning only on the GRM-GPA.
Table 4

Regression Estimates of Differential GPA Inflation by Weighting Approach

<table>
<thead>
<tr>
<th>(A) No Weights</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parent has BA</td>
<td>0.051***</td>
<td>0.010*</td>
<td>0.010*</td>
<td>0.011*</td>
</tr>
<tr>
<td></td>
<td>(0.015)</td>
<td>(0.005)</td>
<td>(0.005)</td>
<td>(0.005)</td>
</tr>
<tr>
<td>GRM Weights</td>
<td>0.467***</td>
<td>0.468***</td>
<td>0.461***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.007)</td>
<td>(0.007)</td>
<td>(0.007)</td>
<td></td>
</tr>
<tr>
<td>Coll. Calculus Grade</td>
<td>-0.003</td>
<td>-0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td>(0.002)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAT-Math</td>
<td></td>
<td></td>
<td></td>
<td>-0.006*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.003)</td>
</tr>
<tr>
<td>R2</td>
<td>.002</td>
<td>.879</td>
<td>.879</td>
<td>.880</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(B) MA Weights</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parent has BA</td>
<td>0.103***</td>
<td>0.059***</td>
<td>0.056***</td>
<td>0.036***</td>
</tr>
<tr>
<td></td>
<td>(0.019)</td>
<td>(0.011)</td>
<td>(0.010)</td>
<td>(0.010)</td>
</tr>
<tr>
<td>GRM Weights</td>
<td>0.513***</td>
<td>0.505***</td>
<td>0.479***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.008)</td>
<td>(0.008)</td>
<td>(0.008)</td>
<td></td>
</tr>
<tr>
<td>Coll. Calculus Grade</td>
<td>0.022***</td>
<td>0.017***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.005)</td>
<td>(0.005)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAT-Math</td>
<td></td>
<td></td>
<td></td>
<td>0.051***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.004)</td>
</tr>
<tr>
<td>R2</td>
<td>.007</td>
<td>.781</td>
<td>.782</td>
<td>.781</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(C) NC Weights</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parent has BA</td>
<td>0.156***</td>
<td>0.108***</td>
<td>0.101***</td>
<td>0.060***</td>
</tr>
<tr>
<td></td>
<td>(0.025)</td>
<td>(0.018)</td>
<td>(0.018)</td>
<td>(0.018)</td>
</tr>
<tr>
<td>GRM Weights</td>
<td>0.558***</td>
<td>0.542***</td>
<td>0.498***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.012)</td>
<td>(0.011)</td>
<td>(0.012)</td>
<td></td>
</tr>
<tr>
<td>Coll. Calculus Grade</td>
<td>0.047***</td>
<td>0.035***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.009)</td>
<td>(0.009)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAT-Math</td>
<td></td>
<td></td>
<td></td>
<td>0.109***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.008)</td>
</tr>
<tr>
<td>R2</td>
<td>.010</td>
<td>.581</td>
<td>.585</td>
<td>.586</td>
</tr>
</tbody>
</table>

Notes. The dependent variable in each regression is a GPA calculated according to a different weighting policy. Panel (A) uses an unweighted GPA, Panel (B) uses Massachusetts’s GPA weighting policy, and Panel (C) uses the weighting policy from North Carolina through the 2014-2015 school year. GRM-weighted GPA, College Calculus Grade, and SAT-M are standardized to have a mean of zero and unit standard deviation. ACT math score scores are transformed to an SAT scale using a concordance. Sample size is 8,336 in all specifications except those in Column (4) because not all students report an SAT or ACT score. Sample size in Column (4) n is 7,021. Heteroskedasticity-robust standard errors, clustered at the college level, in parentheses.

* p < .05, ** p < .01, *** p < .001
4. Concluding Remarks

This article discusses and demonstrates the application of the graded response model (GRM) to the analysis of policies for weighting high school GPAs based on nominal indicators of course difficulty. Using a GRM allows one to avoid the need to rely on a postsecondary criterion to estimate weights, which is a notable advantage given the strong assumptions necessary for criterion-based identification of weights on the high school GPA scale. The GRM is well suited to the task because it uses the same data, retains the key assumptions of conventional GPAs, and relaxes non-essential constraints. For conventional and GRM-adjusted GPAs, the trait of interest is a unidimensional measure of academic achievement, higher grades imply higher trait values, and the trait is estimated from grades received in courses taken. Potential advantages of the GRM are that it can flexibly estimate what conventional GPAs presuppose with respect to grade-by-course difficulty and the reliability with which different courses measure the construct. In the case of GPA adjustment for nominal course difficulty, we show that a GRM-based approach allows one to link letter grade difficulties across different courses to a common scale, which can be transformed to a conventional weighted GPA scale. The ability to compare difficulty across various grade-by-course combinations on a conventional weighted GPA scale would support well-informed policy decisions. We also demonstrate that analyses of differential GPA inflation (i.e. bias) are straightforward to conduct once a GRM-adjusted GPA has been estimated.

The potential use for the GRM—or other latent variable models—for GPA policy analysis extends beyond the study of course rigor. For example, University of California and California State admission policies ignore letter grade plus and minuses when computing GPAs for college admission (UC-CSU, 2016). That information is lost through this policy is obvious, but a less obvious possible consequence is that information loss may be asymmetric, which would occur if
the B grade were much closer to the B- than the B+ (or vice versa). To offer another example, Bulman (2017) found that grades received closer to the end of the high school experience are better predictors of postsecondary outcomes than grades received at the beginning of high school. If GRM-estimated $a_i$ parameters were lower for ninth grade courses compared with tenth-twelfth grade courses, one would be able to support Bulman’s (2017) criterion-based evidence for down-weighting ninth grade courses weights with evidence of greater internal coherence of a GPA that more heavily weights academic performance in later grades. By holding the course content constant and treating courses taken in different years as different items (i.e. “grade9_geometry” and “grade10_geometry”), one could differentiate the role of course content from the time it was taken.

Overall, high school grades play a powerful role in education. They motivate students to study, provide feedback to students about their academic performance, and inform college admission committees about a students’ high school performance. Despite inconsistencies in grading practices across courses, teachers, and schools, grades tend to be predict college success as well as—if not better than—standardized test scores. As a result, many colleges have placed greater emphasis on high school grades in the college admission process. Our study focuses on the applicability of the GRM to the narrow—but important—context of weighted high school GPAs in the college admission context. The absence of a strong empirical foundation for weighted GPA policies is reflected in the wide variety of approaches in use across the United States. A GRM will not ameliorate all flaws in the generation or aggregation of grades, but it can be used to offer sensible, empirical guidance on GPA weighting policies or other policy contexts involving weighted aggregations of ordinal measures.
References


# Table A1

<table>
<thead>
<tr>
<th>Letter Grades Received by Course</th>
<th>Percentage of Letter Grades Received by High School Course</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F</td>
</tr>
<tr>
<td>Algebra1</td>
<td>0.1</td>
</tr>
<tr>
<td>Algebra1-H</td>
<td>0.0</td>
</tr>
<tr>
<td>Algebra2</td>
<td>0.2</td>
</tr>
<tr>
<td>Algebra2-H</td>
<td>0.1</td>
</tr>
<tr>
<td>Geom</td>
<td>0.0</td>
</tr>
<tr>
<td>Geom-H</td>
<td>0.0</td>
</tr>
<tr>
<td>Integrat</td>
<td>0.3</td>
</tr>
<tr>
<td>Integrat-H</td>
<td>0.0</td>
</tr>
<tr>
<td>Oth. Math</td>
<td>0.0</td>
</tr>
<tr>
<td>Oth. Math-H</td>
<td>0.0</td>
</tr>
<tr>
<td>Precalc</td>
<td>0.6</td>
</tr>
<tr>
<td>Precalc-H</td>
<td>0.2</td>
</tr>
<tr>
<td>Statistics</td>
<td>0.3</td>
</tr>
<tr>
<td>Statistics-H</td>
<td>0.0</td>
</tr>
<tr>
<td>Calculus</td>
<td>1.1</td>
</tr>
<tr>
<td>Calculus-H</td>
<td>0.7</td>
</tr>
<tr>
<td>AP CalcAB</td>
<td>0.7</td>
</tr>
<tr>
<td>AP CalcBC</td>
<td>0.0</td>
</tr>
<tr>
<td>AP Stats</td>
<td>0.5</td>
</tr>
<tr>
<td>AP Physics</td>
<td>0.3</td>
</tr>
<tr>
<td>Column %</td>
<td>0.2</td>
</tr>
<tr>
<td>Column n</td>
<td>91</td>
</tr>
</tbody>
</table>