Estimating High School GPA Weighting Parameters with a Graded Response Model

John Hansen
Harvard Graduate School of Education
Cambridge, MA
jdnhansen@gmail.com

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

April, 2018

A revised version of this paper has been accepted for publication at
Educational Measurement: Issues and Practice

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 frequently used 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 high schools. Previous methods of estimating weighting parameters have used regression models with college course performance as the dependent variable. We discuss and demonstrate the suitability of the graded response model for estimating GPA weighting parameters and evaluating traditional weighting schemes.

Key words: Graded response model; item response theory; college admission

The high school grade point average (GPA) is a high-stakes measure of academic achievement. In the college admission process, a small change in GPA can have significant consequences, especially at 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, 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:

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

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

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

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 item response theory (IRT) for examining GPA weighting policies. We used the graded response model (Samejima, 1969) because its underlying assumptions are compatible with conventional GPAs, it uses the same data as conventional GPAs, and graded response model (GRM) 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 IRT applications to course grade data and the theoretical suitability of using a GRM to study high school GPA weighting parameters. In Section 3, we fit a GRM to actual high school course grade data and estimate bonus point parameters for advanced courses on a conventional GPA scale. Section 4 concludes.

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

Many studies of the high school grade point average have shown that high school GPA predicts first-year college grades (Kobrin, Patterson, Shaw, Mattern, & Barbuti, 2008), cumulative college GPA (Geiser & Santelices, 2007), and rates of college graduation (Bowen, Chingos, & McPherson, 2009). Studies have also found that students who take advanced courses in high school, such as AP courses, tend to outperform students who do not (Keng & Dodd, 2008). Geiser and Santelices (2004), Klopfenstein and Thomas (2005), Sadler and Tai (2007b), 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. An unresolved question is the correct methodological approach to estimating how GPAs should be weighted by course type.

Criterion-Based Validation of Weights for Advanced Courses

Previous researchers have relied on an external criterion to estimate GPA weights for advanced courses. Although they reached different conclusions, Geiser and Santelices (2004) and Sadler and Tai (2007a; 2007b) employed similar, criterion-based empirical strategies. They used multiple regression where the outcome was a measure of college achievement, and the independent variables included indicators for participation in advanced coursework. Sadler and Tai (2007b) 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:

$$COL\_CHEM = \beta_0 + \beta_1(HS\_GRADE) + \beta_2(HON\_CHEM) + \beta_3(AP\_CHEM) + \epsilon, \qquad (3)$$

where *COL_CHEM* is the grade in introductory college chemistry, *HON_CHEM* and *AP_CHEM* are dichotomous indicator variables, and standard high school 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, 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:

$$Bonus = \frac{\beta_{HON\_CHEM}}{\beta_{HS\_GRADE}} \tag{4}$$

A fundamental challenge for the Sadler and Tai (2007a; 2007b) methodology is that one must select an external 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 (2007a; 2007b) implicitly assume that college science course performance suffices as a scale for measuring the equivalency between $\beta_{\mathrm{HON\_CHEM}}$ and $\beta_{\mathrm{HS\_GRADE}}$. 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.

An additional 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, 2007b), 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.

Ultimately, criterion and model specification are consequential because many factors are correlated with high school grades, participation in advanced courses, and college outcomes. Different model specifications support different inferences and different interpretations of the high school GPA construct. These challenges motivate the use of item response theory (IRT), because item response models rely only on observed measures of high school performance—the exact data one needs to compute a conventional GPA—and nothing more.

2. Item Response Model GPAs

Statistical methods for adjusting grades received in different courses predate the development of item response theory. Linn (1966) and Young (1993) reviewed methods proposed in the twentieth century, which were almost exclusively non-IRT approaches. Young (1990a) published the first study that used an item response model to link grades from different courses onto a common scale. Studying college performance among undergraduates, he found that the correlation between college GPA and SAT scores increased substantially when using a graded response model-scaled college "GPA" instead of the conventional college GPA. Similar to Young (1990a), recent studies of course-adjusted GPA methods have generally focused on using item response theory to correct the college grade point average or class rank for differential grading stringency across college courses and departments (Bailey, Rosenthal, & Yoon, 2016; Caulkins, Larkey, & Wei, 1996; Johnson, 1997; Lei, Bassiri, & Schultz, 2001; Young, 1990b). An exception is Bassiri and Schultz (2003), who used a Rasch rating scale model to create a universal scale of high school difficulty by using ACT performance as an anchor.

The present study's application of item response theory is distinct from previous work because it focuses explicitly on course-level parameters. In this respect, the present study is perhaps most similar to Korobko, Glas, Bosker, and Luyten (2008), who used IRT to compare the difficulty of university entrance exams in the Netherlands, where students select different

subsets of exams. Our study illustrates item response theory's use for constructing a scale of course difficulty, which could be used to evaluate policies about awarding bonus grade points for advanced course participation. Unlike most previous studies, we did not explicitly seek an estimator of student skill that optimally predicted academic success in college. We chose not to incorporate standardized achievement measures (e.g., SAT or AP exam scores) in our models, because such tests plausibly measure a construct that differs from the one measured by the high school GPA (Willingham, Pollack, & Lewis, 2002). If standardized test scores indeed measure the same construct as the high school GPA (or the construct that one believes the high school GPA ought to measure), including them in the model could improve estimation of grade-by-course difficulty parameters. We omitted them to emphasize that additional data are neither necessary nor necessarily desirable.

The appeal of estimating high school course rigor with item response theory is that one can use the exact same data as conventional GPAs and no additional data, retain the key assumptions of conventional GPAs, and relax non-essential assumptions. Like conventional GPAs, IRT can measure high school achievement in a way that: (1) treats the construct of interest underlying all of a student's grades (hereafter θ) as one-dimensional; (2) assumes that students who receive higher grades in a given course tend to have higher levels of θ than students who receive lower grades in the same course; (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 θ than one who

has a grade in the course. 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 item response theory is that one can relax constraints imposed by 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 θ. 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's statistical properties make it a good candidate for examining the properties of conventional approaches to calculating GPAs and address questions, such as how advanced courses should be weighted. However, similar to other IRT models, the GRM's technical complexity and latent scale make it a less appealing candidate for widespread use as an alternative GPA-weighting algorithm. 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 a latent variable, conventionally assumed to be normally distributed with a mean of zero and unit variance. In estimating parameters, the model finds the θ scale using a cumulative log-odds principle (Ostini & Nering, 2005). Formally, the equation for estimating the GRM can be expressed as:

$$P\left(Y_{ij} \geq k \middle| a_i\, b_{ik}, \theta_j\right) = \frac{1}{1 + \exp\left\{-a_i\left(\theta_j - b_{ik}\right)\right\}}, \tag{5}$$

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 θ; θ 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 θ 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 compare difficulties of letter grades across different courses. True equivalence in difficulty also requires letter grades across the two courses to have identical values of a_i. In this case, fitted probabilities of exceeding the thresholds would be equivalent for all values of θ, 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 θ. If the probability of exceeding thresholds changes little as a function of θ for course i, the lower value of a_i indicates that student grades in course i are less informative measures of θ.

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_{ik}) 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 θ to a 5.0 GPA scale simply by adding three points and truncating GPAs above 5.0 or below zero.

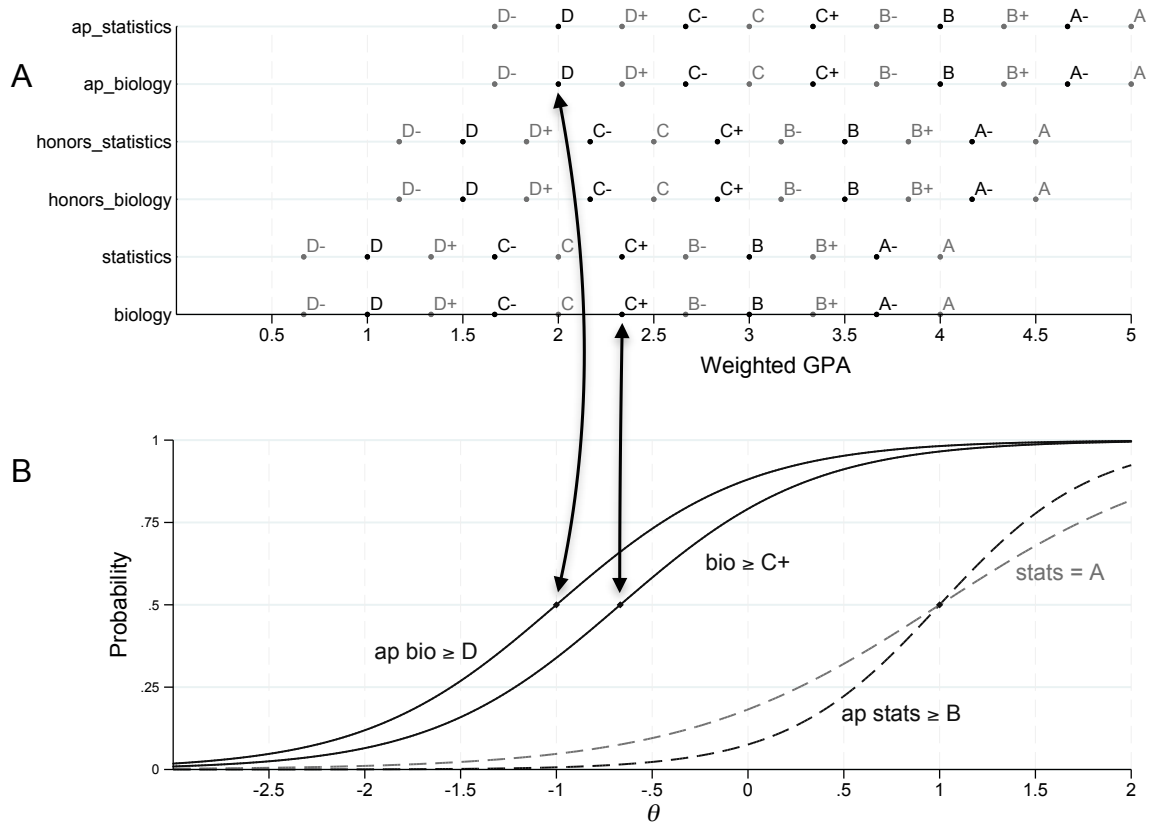

FIGURE 1. Conceptual model for GRM-estimation of high school GPA weighting parameters at the letter-grade-by-course level. (A) Weighted GPA policy in which honors courses receive 0.50 bonus grade points and Advanced Placement courses receive 1.0 bonus grade points. (B) Probability of earning a given course grade or higher as a function of θ.

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 $b_B^{AP\_STATS} = b_A^{STATS}$, the curves intersect where *p*=.50. For student *j* such that $\theta_j = b_B^{AP\_STATS} = b_A^{STATS}$, 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 θ. Nevertheless, the differential is symmetric around *b,* implying equivalent difficulty on average (or, more precisely, equivalent average difficulty for a 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 θ. As a result, for any fixed value of θ, the probability of earning a D or higher in AP Biology is greater than the probability of earning a C+ in standard biology.

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

$$P\left(Y_{ij} \geq k \middle| a_i\, b_{ik}, \theta_j\right) = \frac{1}{1 + \exp\{-a_i(\theta_j - b_{ik})\}} = \frac{1}{1 + \exp\{\frac{-a_i}{m}(m(\theta_j - b_{ik}))\}} \tag{6}$$

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 θ-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 answer to the "bad question" of whether an item response model fits is, in binary terms, 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 θ. A GRM does not ameliorate all shortcomings of grades or conventional GPA aggregation approaches. Many potential 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 what one believes student 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 here. Our key point is that a GRM approach is well-suited to estimating how various grade-by-course combinations can be compared to one another on a common scale.

Overall, using a GRM to estimate letter grade-by-course difficulty parameters on a latent scale can, at an abstract level, be characterized as a linking study with non-equivalent groups, a common-item design, and concurrent calibration (Kolen & Brennan, 2004). Simultaneous estimation of course parameters (concurrent calibration) requires the existence of multiple courses and crossing of students and courses. Courses with the same name are treated as common items, allowing all courses and students to be linked to a common scale. While courses taught at different schools are obviously non-equivalent in many respects, weighted GPA policies regularly treat these courses equivalently. Consequently, differences in course difficulty associated with nominal course identifiers are the policy-relevant quantities of interest. Interpretation of b_{ik} parameters where high school courses are the "items" is analogous to canonical GRM applications under a few conditions, which are discussed in the appendix (S1).

3. Demonstration

In this section, we fit a GRM to high school course grade data and use the results to estimate bonus point (i.e., weighting) parameters for advanced courses on a conventional GPA scale.

3.1 Data and the GRM Model

The data for this section of the study are from the Factors Influencing College Success in Mathematics survey (FICSMath), which was gathered in a very similar fashion as the data used by Sadler and Tai (2007a; 2007b). The FICSMath survey was administered to 10,492 first-year college calculus students from 135 randomly selected U.S. colleges during the 2009-2010 academic year. The sample was intended to be approximately—but not precisely—representative of first-year college calculus students. The sample included public, private, large, small, selective, non-selective, religious, and non-religious institutions from across the United States. Twenty-eight institutions had "Community College" in their name, and six institutions were classified as "Most Competitive" in *Barron's Profiles of American Colleges* (2009).

Students took the FICSMath survey at the beginning of the course. At the end of the term, the calculus instructor recorded each student's grade from the course. Most of the survey items focused on pedagogical decisions made by teachers that could contribute to college success in mathematics and science. It also included items about the student's high school, the student's demographic background, and, most importantly for this study, the mathematics courses they took in high school and the letter grades they received in those courses. We focused only on quantitative courses (mathematics courses and AP Physics), because the survey has limited data on non-quantitative coursework and our model assumes a unidimensional construct.

We fit the graded response model using the "irt grm" command in Stata 14 (StataCorp, 2015). The default maximum likelihood estimator approximates integrals with mean and variance

adaptive Gauss-Hermite quadrature. Initial attempts to fit the model encountered some convergence challenges. In response, we collapsed D- and D grades into a single category, because D- grades were relatively rare in our sample (A1). At this point, the algorithm converged within two minutes without issuing errors or warnings. The 853 students who did not report any course grades were excluded from estimation. Next, we estimated the posterior mean of θ for each student using empirical Bayes. This is the GRM analog to GPA.

Next, to create a consistent sample for subsequent analyses, in this order, we listwise deleted 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). The primary goal of these restrictions was to create a sample similar to Sadler and Tai (2007a; 200b), who included an external criterion and demographic covariates in their models. Table 1 shows that the students taking calculus as first-year college students were relatively high-performing high school students. The average unweighted GPA for quantitative courses in the analytical sample was 3.50.

TABLE 1
Descriptive Statistics for the Analytical Sample

	Mean	Min	P25	Median	P75	Max
Sample (N = 8336)						
Calculus Grade	79.79	12.00	74.50	84.50	92.00	100.00
GPA: GRM Weights	0.03	-2.94	-0.60	0.00	0.81	1.71
GPA: No Weights	3.50	0.75	3.20	3.61	3.93	4.00
GPA: Weighted (1.0 AP Bonus)	3.73	0.75	3.40	3.84	4.13	5.00
Total Course Grades Reported	4.77	3.00	4.00	5.00	5.00	12.00
Hon. Course Grades Reported	1.43	0.00	0.00	0.00	3.00	8.00
AP Course Grades Reported	0.49	0.00	0.00	0.00	1.00	4.00
Parent has Bachelor's	0.63	0.00	0.00	1.00	1.00	1.00
SAT-Math*	612.75	200	560	620	680	800

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.
*SAT-mathematics also includes ACT Math scores translated to an SAT scale using an SAT/ACT concordance. SAT and ACT scores were missing for 1,315 students.

Table 2 shows correlations among variables from Table 1. Similar to previous studies, the IRT estimate of student academic skill, labeled "GRM Weights," had a stronger correlation with the relevant criterion outcome than the unweighted GPA. The correlation was .352 (CI [.333, .370]), and its 95% confidence interval excluded .326, the unweighted GPA's correlation coefficient. The GRM-GPA's confidence interval did not exclude the conventionally weighted GPA's (1.0 bonus for AP) correlation coefficient, .347. At the same time, parental educational attainment, a measure of student socioeconomic status (SES), was positively correlated with calculus grade (r =.085) and more strongly correlated with the conventionally weighted GPA (r = .086, CI [.064, .107]) than the GRM-weighted GPA (r =.042, CI[.02, .063]). These covariance patterns illustrate the challenges of using a criterion-based approach to validate GPA weights for advanced courses. Overall, stronger correlations with a criterion, compared with conventional GPAs and weaker correlations with other variables (e.g., parental education) help establish credibility for the model's estimates of grade-by-course difficulty, but that is all.

TABLE 2
Correlation Matrix

	Calc. Grade	GRM GPA	GPA	W. GPA	Total Course	Hon. Course	AP Course	Parent BA	SAT Math
Calculus Grade	1.000								
GPA: GRM Weights	0.352	1.000							
GPA: No Weights	0.326	0.937	1.000						
GPA: Weighted	0.347	0.882	0.913	1.000					
Total Course Grades	0.157	0.210	0.174	0.279	1.000				
Hon. Course Grades	0.130	0.176	0.144	0.469	0.279	1.000			
AP Course Grades	0.145	0.151	0.088	0.341	0.596	0.259	1.000		
Parent has Bachelor's	0.085	0.042	0.049	0.086	0.085	0.102	0.070	1.000	
SAT-Math*	0.270	0.339	0.308	0.387	0.263	0.237	0.244	0.172	1.000

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. All correlations are statistically significant ($\alpha < .001$). *SAT-Math includes ACT Math scores translated to an SAT scale using an SAT/ACT concordance. SAT and ACT scores were missing for 1,315 students.

Figure 2 plots the GRM estimates of grade-by-course difficulty (b_{ik}) on the θ scale. For this sample, which was restricted to students enrolling in first-year college calculus, the ranking of

difficulty parameter estimates generally supported the practice of GPA weights for AP courses—particularly for AP Calculus. The lower rate at which students with high grades in other classes received A's in AP Calculus is the GRM's empirical basis for identifying the course as particularly difficult. The proportion of A's received was lowest in AP Calculus BC and second lowest in AP Calculus AB (Table A1). Students who reported taking either AP Calculus course had higher average grades in their other courses (M = 3.65, SD = .40) than students who did not take AP calculus (M = 3.47 5, SD = .52), $t(8334) = 15.14$, $p < .001$).

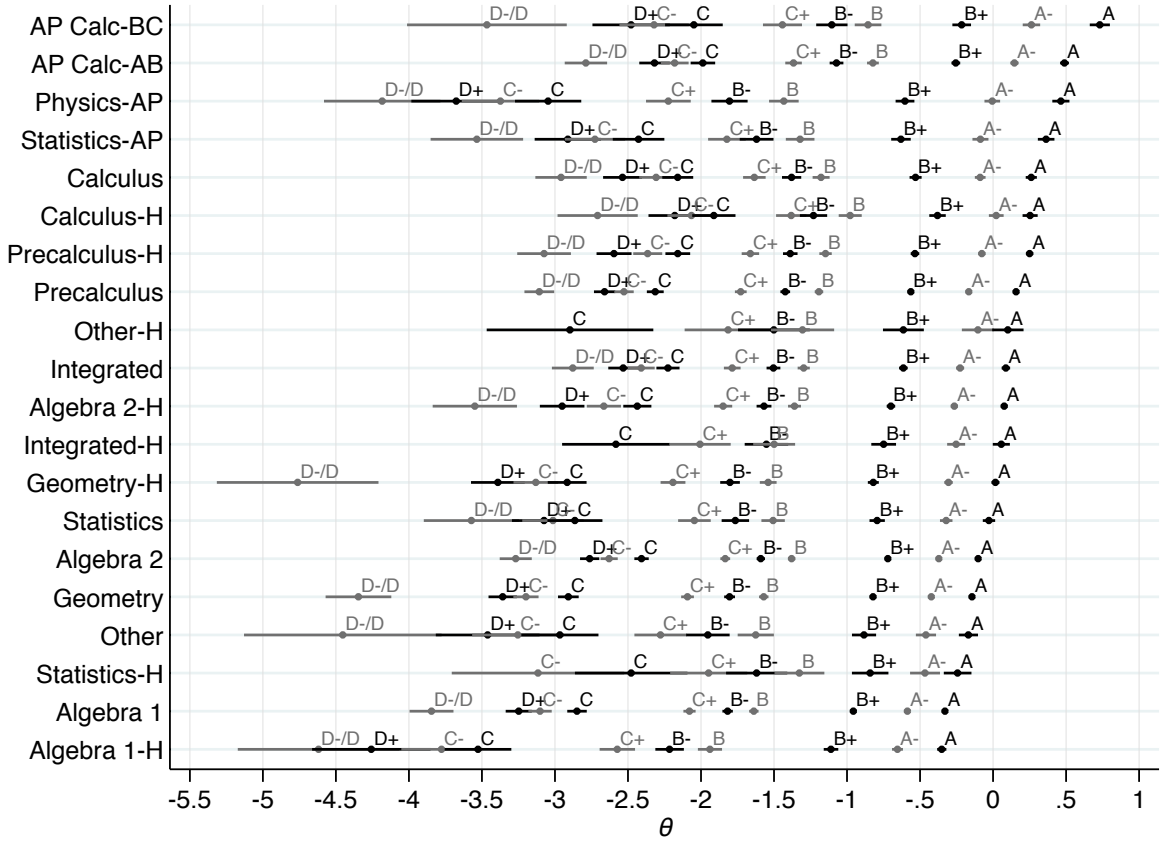

FIGURE 2. Each point indicates the mathematics skill (θ) necessary to have a 0.50 probability of earning the given letter grade or higher in the given course. Confidence intervals are ± 1 SE. Grades of D- and D are collapsed into a single category due to sparse data at the low end of the grade distribution (Table A1).

Overall, the results presented in Figure 2 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 students with θ around -0.25). In most cases, it appears that approximately 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.

3.2 Linking Latent-scale Parameters to a GPA Scale

The $BONUS_i$ parameters of interest from Equation (2) are not directly estimated by the GRM, but the GRM's θ-scale estimates of grade-by-course difficulty, $\hat{b}_{ik}^{\theta}$, can be transformed to a conventional GPA scale and then used to estimate $BONUS_i$ parameters. We used a linear linking (Kolen & Brennan, 2004) approach to identify the linear transformation that mapped the $\hat{b}_{ik}^{\theta}$ for standard courses to a conventional 4.0 GPA, and applied the same linear transformation to honors and advanced courses. Figure 3 displays the results of the transformation: a linear shift of the θ-scale estimates from Figure 2 onto a GPA scale (technical details in appendix S2).

The difference between an unweighted grade point value (4.0 grade points for an A in AP Statistics) and the estimated grade point value (4.25 grade points for an A in AP Statistics) is our estimate for $BONUS_{AP\_STATS}$. The largest estimated bonus point adjustment was 0.55 grade points for AP Calculus BC. One could use the results in Figure 3 to design a weighting system where each course was weighted differently, or one could use the average $BONUS_i$ for AP courses to design a weighting system where all AP courses receive the same adjustment, $BONUS_{AP}$. We estimated that the average GPA-scale difference in difficulty between standard and AP courses, $BONUS_{AP}$, was 0.25 letter grade points—a bit shy of the difference between an A and an A-. Our estimate of $BONUS_{HON}$ was 0.02 letter grade points—virtually no difference.

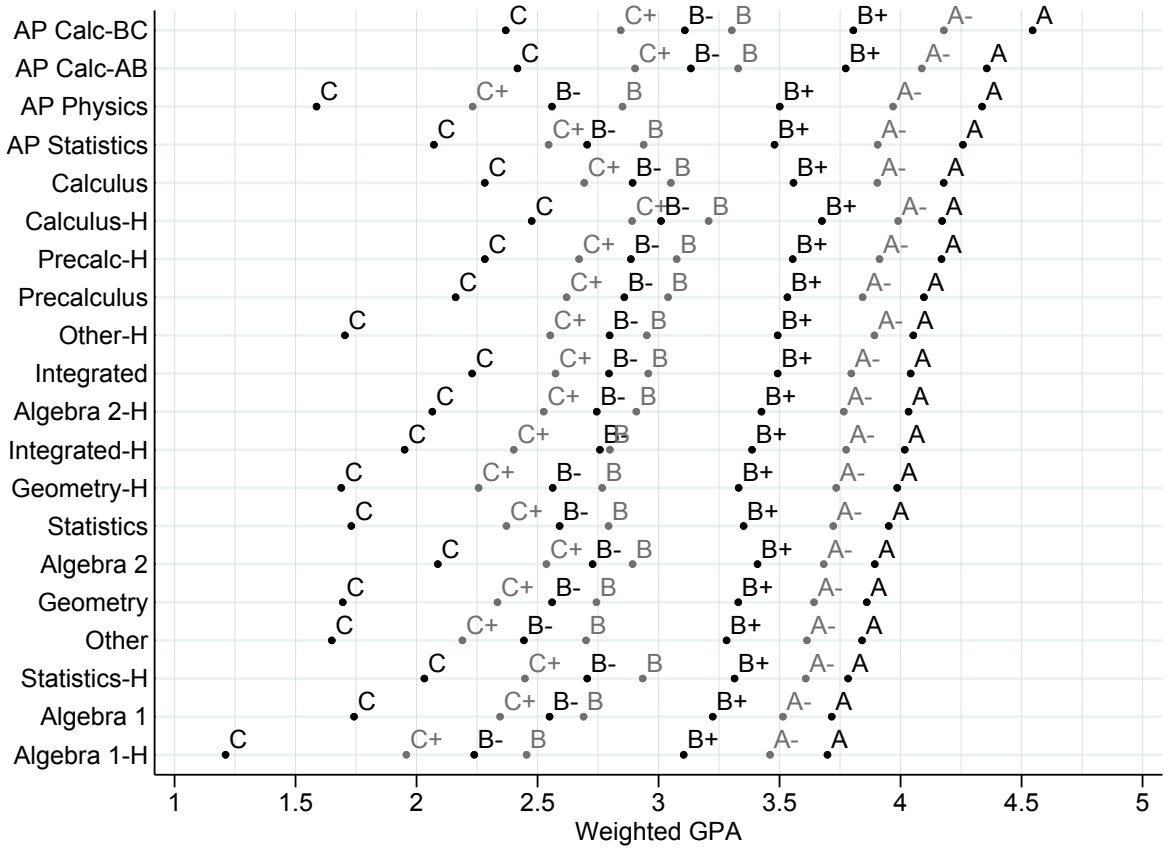

FIGURE 3. Point estimates of grade by course difficulty on a weighted high school GPA scale. Grades below C were estimated imprecisely due to sparse data, and were omitted from the linking process (see S2 for details).

3.3 Limitations

The purpose of this study is to discuss and demonstrate the GRM approach to evaluating policies for weighting high school GPAs, not to make a definitive claim about the "correct" high school GPA weighting parameters across all U.S. schools. Our study included only one IRT method for estimating latent scale parameters, only one method for transforming latent scale parameters to GPA scale parameters, and only one method for aggregating $BONUS_i$ into $BONUS_{AP}$. Our decisions prioritized clarity in illustration of the methodology over other factors. How such decisions impact estimates of relevant parameters merits further study. Another potential limitation is that high school grades in our sample were restricted to quantitative

coursework for students who chose to take college calculus, a relatively high-achieving sample. An appealing feature of estimating high school GPA parameters with an item response model is that, theoretically, if the model fits the data, parameter estimates will be sample-invariant (up to a linear transformation). Future research should explore the extent to which parameter estimates are indeed sample-invariant by comparing estimates across heterogeneous samples.

4. Concluding Remarks

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 students' high school performance. Despite inconsistencies in grading practices across courses, teachers, and schools, grades tend to predict college success as well as—if not better than—standardized test scores do. As a result, in recent years, many colleges have placed greater emphasis on high school grades in the college admission process.

To account for differences in grading standards between standard and advanced courses, GPAs are often adjusted to account for nominal indicators of course rigor, such as "Advanced Placement." This study discussed and demonstrated how item response theory could be used to estimate course difficulty differentials between courses. The ability to compare the difficulty of various grade-by-course combinations on a conventional weighted GPA scale can support well-informed policy decisions for weighting high school GPAs.

References

Arrow, K. J. (1951). *Social choice and individual welfare*. Cowles Commission Monograph 12. New York: John Wiley and Sons.

Atkinson, R. C., & Geiser, S. (2009). Reflections on a century of college admissions tests. *Educational Researcher*, 38, 665–676.

Bailey, M. A., Rosenthal, J. S., & Yoon, A. H. (2016). Grades and incentives: assessing competing grade point average measures and postgraduate outcomes. *Studies in Higher Education*, *41*(9), 1548-1562.

Barron's Profiles of American Colleges (2009) *Barron's Profiles of American Colleges, 2009.* Hauppauge, NY: Barron's Educational Series Inc.

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

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.

Caulkins, J. P., Larkey, P. D., & Wei, J. (1996). Adjusting GPA to reflect course difficulty. Retrieved from http://repository.cmu.edu/cgi/viewcontent.cgi?article=1042&context=heinzworks

Cal-Grant. (2016, Sept. 21). Worksheet to calculate Cal-Grant GPA. Retrieved from http://www.oakparkusd.org/cms/lib5/CA01000794/Centricity/Domain/165/Cal-Grant_GPA_Calculation.pdf

Clinedinst, M., Koranteng, A.-M., & Nicola, T. (2016). *State of college admission.* Arlington, VA: National Association for College Admission Counseling. Retrieved from https://indd.adobe.com/view/c555ca95-5bef-44f6-9a9b-6325942ff7cb

Geiser, S., & Santelices, M. V. (2004). The role of Advanced Placement and honors courses in college admissions. (Research & Occasional Paper Series: CSHE.4.04.) Berkeley: University of California Center for Studies in Higher Education. Retrieved from http://www.cshe.berkeley.edu/sites/default/files/shared/publications/docs/ROP.Geiser.4.04.pdf

Geiser, S., & Santelices, M. V. (2007). Validity of high-school grades in predicting student success beyond the freshman year. (Research & Occasional Paper Series: CSHE.6.07.) Berkeley: University of California Center for Studies in Higher Education. Retrieved from http://files.eric.ed.gov/fulltext/ED502858.pdf

Hiss, W. C., & Franks, V. W. (2014). Defining promise: Optional standardized testing policies in American college and university admissions. *Report of the National Association for College Admission Counseling (NACAC).* Retrieved from http://www.nacacnet.org/research/research-data/nacac-research/Documents/DefiningPromise.pdf

Horn, C. L., & Flores, S. M. (2003). Executive summary: Percent plans in college admissions. A comparative analysis of three states' experiences. Retrieved from https://civilrightsproject.ucla.edu/research/college-access/admissions/percent-plans-in-college-admissions-a-comparative-analysis-of-three-states2019-experiences/

Johnson, V.E. (1997). An Alternative to Traditional GPA for Evaluating Student Performance. *Statistical Science, 12,* 215-269.

Keng, L., & Dodd, B. G. (2008). *A comparison of college performances of AP and non-AP student groups in 10 subject areas*. New York: College Board.

Klopfenstein, K., & Thomas, M. K. (2005). The advanced placement performance advantage: Fact or fiction. *Texas Christian University and Mississippi State University*.

Kobrin, J. L., Patterson, B. F., Shaw, E. J., Mattern, K. D., & Barbuti, S. M. (2008). *Validity of the SAT for Predicting First-Year College Grade Point Average*. Research Report No. 2008-5. New York: College Board.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking*. New York: Springer.

Korobko, O. B., Glas, C. A., Bosker, R. J., & Luyten, J. W. (2008). Comparing the difficulty of examination subjects with item response theory. *Journal of Educational Measurement*, *45*(2), 139-157.

Lang, D. M. (2007). Class rank, GPA, and valedictorians: How high schools rank students. *American Secondary Education*, *35*(2), 36.

Lei, P. W., Bassiri, D., & Schultz, E. M. (2001). *Alternatives to the Grade Point Average as a Measure of Academic Achievement in College*. (Report Np. TM033669). ACT Research Report Series. American College Testing Program. Iowa City, Iowa.

Linn, R. L. (1966). Grade adjustments for prediction of academic performance: A review. *Journal of Educational Measurement, 3*(4), 313-329.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Routledge.

Massachusetts Department of Higher Education. (2013). *Admission standards for the Massachusetts State University System and the University of Massachusetts. Guidance*

for high school guidance counselors. Retrieved from http://www.mass.edu/shared/documents/admissions/admissionsstandards.pdf

Miami-Dade County Public Schools. (2016) *Rank in class—grade point average.* School Board rule 6Gx13-5B-1.061. Retrieved from: http://www.dadeschools.net/schoolboard/rules/Chapt5/5b-1.061.pdf

North Carolina State Board of Education Policy Manual (2015). Policy outlining standards to be incorporated into the electronically generated high school transcript. GS 116-11(10a). Retrieved from http://sbepolicy.dpi.state.nc.us/policies/GCS-L-004.asp?pri=01&cat=L&pol=004&acr=GCS

Ostini, R., & Nering, M. L. (2005). *Polytomous item response theory models*. Thousand Oaks: Sage.

Rothstein, J. M. (2004). College performance predictions and the SAT. *Journal of Econometrics*, *121*(1), 297-317.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581-592.

Sadler, P. M., & Tai, R. H. (2007a). Accounting for advanced high school coursework in college admission decisions. *College and University, 82(4)*, 7.

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

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*.

Sawyer, R. (2013) Beyond correlations: Usefulness of high school GPA and test scores in making college admissions decisions. *Applied Measurement in Education*, *26*(2), 89-112, DOI: 10.1080/08957347.2013.765433.

StataCorp (2015). *Stata Statistical Software: Release 14.* College Station, TX: StataCorp LP.

Stricker, L. J., Rock, D. A., Burton, N. W., Muraki, E., & Jirele, T. J. (1994). Adjusting college grade point average criteria for variations in grading standards: A comparison of methods. *Journal of Applied Psychology*, *79*(2), 178.

Thissen, D. (2016). Bad questions: An essay involving item response theory. *Journal of Educational and Behavioral Statistics*, *41*(1), 81-89.

University of California. (2016, Sept. 21). Calculating the UC/CSU GPA. Retrieved from http://collegetools.berkeley.edu/documents/cat_113-128/Calculating_GPA.pdf

Vickers, J. M. (2000). Justice and truth in grades and their averages. *Research in Higher Education*, *41*(2), 141-164.

Warne, R. T. (2017). Research on the academic benefits of the Advanced Placement program: Taking stock and looking forward. *SAGE Open, 7*(1).

Warne, R. T., Nagaishi, C., Slade, M. K., Hermesmeyer, P., & Peck, E. K. (2014). Comparing weighted and unweighted grade point averages in predicting college success of diverse and low-income college students. *NASSP Bulletin*, *98*(4), 261-279.

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.

Young, J. W. (1990a). Adjusting the cumulative GPA using item response theory. *Journal of Educational Measurement*, 27, 175– 186.

Young, J. W. (1990b). Are validity coefficients understated due to correctable defects in the GPA? *Research in Higher Education*, *31*(4), 319-325.

Young, J. W. (1993). Grade adjustment methods. *Review of Educational Research*, *63*(2), 151-165.

Zimmerman, S. D. (2014). The returns to college admission for academically marginal students. *Journal of Labor Economics*, *32*(4), 711-754.

Supplementary Materials

S1. Interpreting b_{ik} Parameters When High School Courses Are Items

The GRM parameters of interest for course rigor analysis are b_{ik}, and their interpretation is analogous to typical GRM parameters (i.e., for standardized test items) under the following conditions: (1) Grade by course difficulty parameters measure grading stringency net of course-specific differences in student effort, and (2) θ is treated as time-dependent.

Interpreting b_{ik} parameters as measures of grading stringency net of course-specific student effort is not inconsistent with typical GRM parameter interpretation. 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 courses (or 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 θ 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 θ, 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 θ—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 period of several years. Time is not explicitly modeled, and Y_{ij} is theoretically a function of θ_j, item parameters, and *when students took the course*. A student who excelled in calculus at 12th grade would not necessarily have fared well in ninth grade, before they had studied advanced algebra and trigonometry. One resolution is to reinterpret θ 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' θ 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 12th grade out of reach). It would presume that a student who took precalculus in 11th grade and chose not to take calculus the following year can be compared in terms of θ to a student who chose to take AP calculus in 12th grade. Ultimately, while this interpretation departs from canonical GRM applications, it is consistent with conventional interpretations of the GPA construct.

S2. Technical Details for Linear Linking of Parameters from the θ Scale to the GPA Scale

We use a simple linear linking (Kolen & Brennan, 2004) approach to transform course difficulty parameters from a θ scale to a conventional GPA scale. More sophisticated approaches may be preferable. Our linear equation for linking b_{ik}^{θ} to the GPA scale is $b_{ik}^{GPA} = mb_{ik}^{\theta} + z$, where *m* denotes the variance rescaling parameter and *z* denote the mean re-centering parameter. In this case, we have one set of parameter estimates for all courses on the θ scale, $\hat{b}_{ik}^{\theta}$, and one set of known, policy-dictated parameters for letter grade point values in standard courses, $b_{k}^{GPA^*}$ (the $b_{k}^{GPA^*}$ vector contains the standard course grade point values plotted in Panel A of Figure 1). This allows one to use an estimator of *m* that exploits the known, fixed distance between letter grade values in standard courses on the target scale.

$$m = \frac{b_{K2}^{GPA^*} - b_{K1}^{GPA^*}}{\frac{1}{n}\sum_{i=1}^{n} b_{iK2}^{\theta} - b_{iK1}^{\theta}}, \tag{7}$$

where $b_{K2}^{GPA^*}$ and $b_{K1}^{GPA^*}$ are two fixed points on the known GPA scale, and b_{iK2}^{θ} and b_{iK1}^{θ} are the corresponding points on the θ scale for standard course *i.* In our sample, b_{ik} parameters are estimated imprecisely at the lower end of the distribution, so we estimate *m* using the distance between C and A grades (b_{ik} estimates for lower grades are also more sensitive to a_i, as shown in Figure A2). In this case, the numerator of Equation (7) would be $b_A^{GPA^*} - b_C^{GPA^*}$, 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 θ scale is compressed (or stretched) to match the conventional GPA scale. For our sample, Figure A1 shows the compression of points in Panel B compared to Panel A. Note that Equation (7) only depends on two fixed points, an upper and lower bound. Parameter estimates between A and C can be ignored in this estimator because the distance between A and C is equivalent to the sum of intermediate distances (i.e., (A-C) = (A-B) + (B-C)).

We estimate *z* with the mean difference in b_{ik} across the two scales for standard courses, after adjusting b_{ik}^{θ} by a factor of *m*. Continuing to use b_{ik} for C through A as our scale anchors, the equation is:

$$z = \frac{1}{nq} \sum_{i \in I} \sum_{k \in K} b_{ik}^{GPA^*} - m b_{ik}^{\theta}, \tag{8}$$

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*.

Figure A1 provides a visual illustration of the scale transformation for calculus and statistics. Panel A plots $\hat{b}_{ik}^{\theta}$, and Panel B plots $\hat{b}_{ik}^{GPA}$. If the $\hat{b}_{ik}$ for AP and standard courses were the same—that is, if AP and standard courses were estimated to be equally difficult—they would fall on the main diagonal in both panels. The "1.0 Point Bonus Line" in Panel B indicates where the

$\hat{b}_{ik}$ would lie if we had found that Advanced Placement courses in statistics and calculus were exactly one letter grade 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 average vertical distance between the letter grades labeled "stats" and the main diagonal in Panel B.

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 ($\bar{b}^{GPA}_{AP\_STATS} - \bar{b}^{GPA}_{STATS}$), or the difference in difficulty between an AP course and the average standard course ($\bar{b}^{GPA}_{AP\_STATS} - \bar{b}^{GPA}_{I}$)—or perhaps something else—is ultimately a policy question. Unweighted or weighted average differences in $\hat{b}^{GPA}_{ik}$ and the comparison course(s) are potential *BONUS* estimators. Possible candidates for weights could be the number of students taking the course in the population of interest or a function of the standard error for $\hat{b}^{GPA}_{ik}$.

In our study, we estimated the unweighted average difference in b^{GPA}_{ik} 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:

$$BONUS_{AP} = \frac{1}{cq}\sum_{p\in P}\sum_{k\in K} b^{GPA}_{pk} - \frac{1}{nq}\sum_{i\in I}\sum_{k\in K} b^{GPA}_{ik}, \tag{9}$$

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 $BONUS_{AP}$ was 0.25 letter grade points, and the analogous estimate for $BONUS_{HON}$ was 0.02 letter grade points.

Table A1

Letter Grades Received by Course

	Percentage of Letter Grades Received by High School Course											
	F	D-/D	D+	C-	C	C+	B-	B	B+	A-	A	*Total*
Algebra1	0.1	0.3	0.2	0.4	3.4	2.3	2.2	13.2	10.3	8.2	59.3	5314
Algebra1-H	0.0	0.0	0.1	0.1	1.4	1.4	1.8	10.5	9.9	8.3	66.3	2390
Algebra2	0.2	0.6	0.4	0.8	3.9	2.9	3.5	16.8	11.4	9.2	50.3	4999
Algebra2-H	0.1	0.2	0.3	0.4	2.4	2.2	2.7	13.4	13.0	11.6	53.6	2923
Geom	0.0	0.5	0.2	0.6	4.1	2.9	3.2	16.1	11.3	8.7	52.4	5144
Geom-H	0.0	0.4	0.3	0.4	2.5	2.9	2.9	12.7	13.3	9.8	54.9	2822
Integrat	0.3	0.7	0.3	0.7	3.2	3.2	3.5	17.5	13.8	11.5	45.3	1370
Integrat-H	0.0	0.0	0.4	0.0	1.2	2.3	0.4	9.3	14.3	10.8	61.4	259
Oth. Math	0.0	0.4	0.2	0.8	3.3	2.9	3.7	14.0	10.3	7.6	56.7	485
Oth. Math-H	0.0	0.0	0.0	0.6	5.2	3.5	2.9	13.8	13.8	5.8	54.6	174
Precalc	0.6	0.9	0.5	1.0	4.8	4.3	4.4	16.4	12.4	10.3	44.5	4282
Precalc-H	0.2	0.5	0.5	0.6	2.9	3.1	3.8	14.4	14.5	10.7	48.8	2554
Statistics	0.3	0.4	0.1	0.2	3.5	2.4	3.3	14.3	13.4	9.1	52.9	911
Statistics-H	0.0	0.0	0.0	1.0	2.4	2.9	3.9	8.7	8.7	6.3	66.0	206
Calculus	1.1	1.0	0.9	0.8	4.2	3.6	3.6	14.9	12.8	10.7	46.5	1291
Calculus-H	0.7	1.4	0.5	0.9	4.7	2.2	4.3	13.7	13.0	7.9	50.5	554
AP CalcAB	0.7	1.1	0.5	0.9	5.8	4.4	5.0	15.4	12.5	10.9	42.8	2032
AP CalcBC	0.0	1.2	0.5	0.9	4.7	5.0	4.7	17.5	16.0	13.7	35.9	424
AP Stats	0.5	1.2	0.4	1.4	4.7	2.3	3.9	13.6	14.8	12.2	44.9	770
AP Physics	0.3	0.3	0.3	0.7	3.7	3.5	4.4	16.1	15.2	12.4	43.1	870
Column %	0.2	0.6	0.3	0.6	3.6	3.0	3.3	14.7	12.2	9.7	51.8	100%
Column n	91	219	130	251	1435	1178	1317	5846	4835	3869	20603	39774

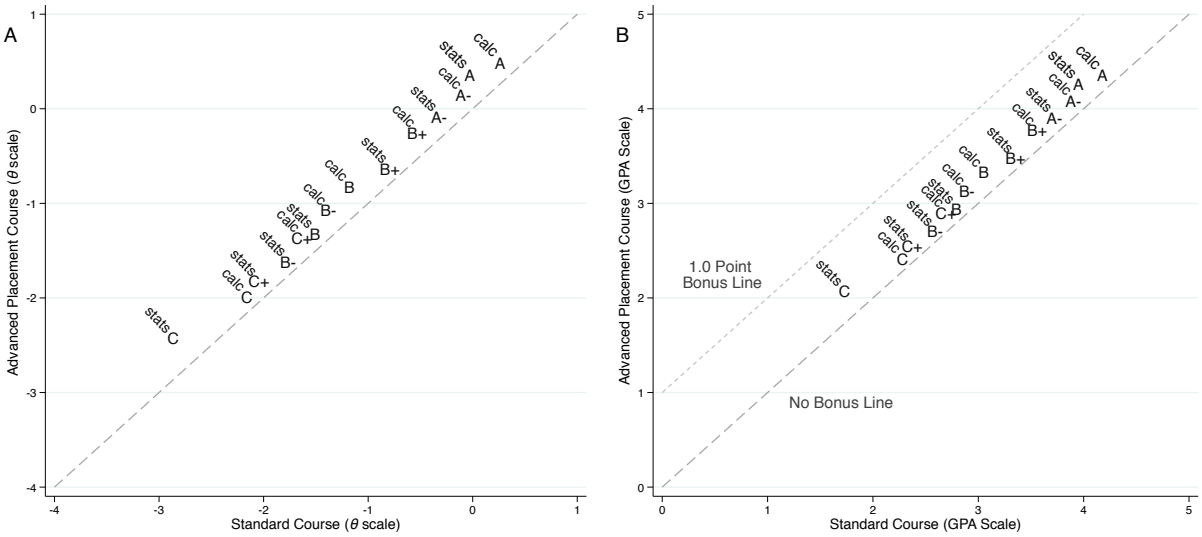

FIGURE A1. Comparison across scales of estimated b_{ik} parameters in calculus and statistics for standard and Advanced Placement courses. For calculus, AP Calculus AB is plotted. (A) θ scale. (B) GPA scale.

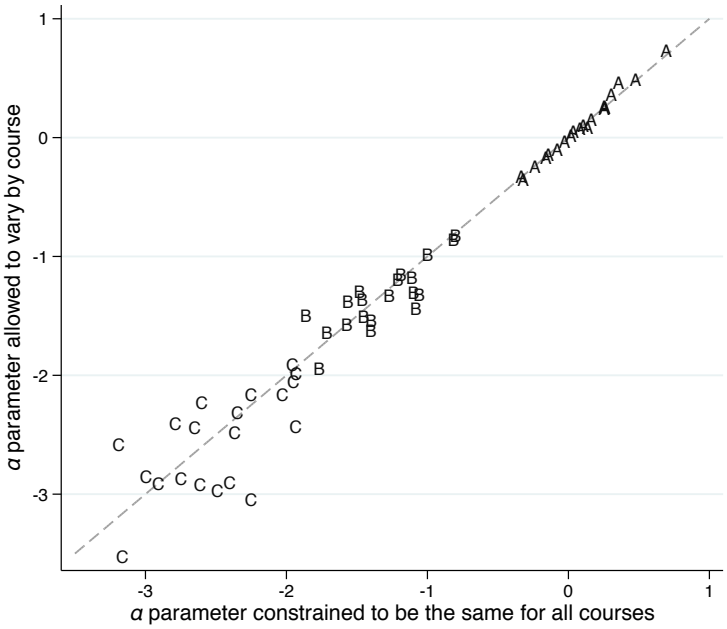

FIGURE A2. Comparison of b_{ik} parameters estimated from a graded response model in which *a* parameters are allowed to vary by course (y axis) and a graded response model in which *a* parameters are constrained to be equal for all courses (x axis).