EdWorkingPaper No. 24-1109

Making the Grade: Accounting for Course Selection in High School Transcripts with Item Response Theory

Kenneth A. Shores
University of Delaware

Sanford R. Student
University of Delaware

We use student-level administrative data from Delaware for 43,767 high school students across five 12th grade cohorts from 2017 to 2021. We apply Item Response Theory (IRT) to high school transcript data, treating courses as items and grades as ordered responses, to estimate both student transcript strength (θ) and course difficulty. We prove, via construct and predictive validation and simulation, that θ improves upon GPA because it accounts for ability selection into courses with variable difficulty. Compared to the SAT, θ shows smaller racial/ethnic gaps but substantially larger gender gaps that indicate boys underperform in their courses relative to their standardized test scores. We conclude by discussing significant methodological—such as grade-inflation and cross-school heterogeneity in course offerings—and practical challenges that remain before such measures could be considered for high-stakes applications.

VERSION: December 2024

Suggested citation: Shores, Kenneth A., and Sanford R. Student. (2024). Making the Grade: Accounting for Course Selection in High School Transcripts with Item Response Theory. (EdWorkingPaper: 24 -1109). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/48d6-mx29

Making the Grade: Accounting for Course Selection in High School Transcripts with Item Response Theory

Kenneth Shores

Sanford R. Student

Abstract

We use student-level administrative data from Delaware for 43,767 high school students across five 12th grade cohorts from 2017 to 2021. We apply Item Response Theory (IRT) to high school transcript data, treating courses as items and grades as ordered responses, to estimate both student transcript strength ($\hat{\theta}$) and course difficulty. We prove, via construct and predictive validation and simulation, that $\hat{\theta}$ improves upon GPA because it accounts for ability selection into courses with variable difficulty. Compared to the SAT, $\hat{\theta}$ shows smaller racial/ethnic gaps but substantially larger gender gaps that indicate boys underperform in their courses relative to their standardized test scores. We conclude by discussing significant methodological—such as grade-inflation and cross-school heterogeneity in course offerings—and practical challenges that remain before such measures could be considered for high-stakes applications.

*Kenneth Shores is associate professor at the University of Delaware in the School of Education and the Biden School of Public Policy and Administration. Sanford Student (srstu@udel.edu) is assistant professor at the University of Delaware in the School of Education. The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through R305S210008 to the Delaware Department of Education. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. Data and programs for replication will be made available at https://www.openicpsr.org. The replication files will not include state administrative data, which are available for researchers by request through the Delaware Department of Education.

Introduction

Recent shifts toward test-optional college admissions policies have intensified debate about how to evaluate student preparedness for college (e.g., Dessein, et al., 2025). While proponents argue these policies increase equity and access (Bennett, 2021; Belasco, et al., 2015), critics contend that standardized tests provide valuable information about student ability (Dessein, et al., 2025) or that alternative measures of college readiness presented to admissions officers may be less equitable than the test scores themselves (Newman et al., 2022; Rothstein, 2022). This debate raises important questions: what information about student ability is contained in high school transcripts, how does this information compare to standardized test scores, and how are inferences regarding inequality affected by the measurement tool?

These questions connect to a broader methodological challenge in educational measurement. Grade point average (GPA), the standard summary of transcript data, is almost certainly an imperfect means of comparing students' academic performance in high school when students select into courses based on course difficulty and their anticipated performance in harder or easier courses. While a student taking advanced courses might earn the same GPA as a peer in standard courses, these identical GPAs likely reflect different levels of transcript strength. This suggests that transcript data might contain more information about college readiness than is captured by GPA alone, yet researchers frequently use GPA scores as both predictors and dependent variables in various applications (Backes et al., 2024; Jackson, 2018).

We address both the practical and methodological aspects of this challenge. First, we apply Item Response Theory (IRT; Hambleton & Swaminathan, 1985) to high school transcript data, treating courses as items and grades as ordered responses to estimate both student ability and course difficulty. This approach explicitly accounts for the differential difficulty of courses in producing scores, potentially improving upon GPA as a measure of transcript strength and college readiness. Second, we compare our IRT-based measure of transcript strength ($\hat{\theta}$) to other measures of student ability used in both practice and by researchers, student GPA and SAT. We demonstrate construct and predictive validity for this IRT-based measure and conclude by characterizing how student subgroup differences appear using SAT scores and transcript data.

Our results provide two main contributions. For policy discussions around college admissions, we offer new evidence about the information contained in transcript data that admissions officers observe, including how different measures characterize educational inequality across student subgroups. For researchers, we demonstrate that IRT estimation can substantially improve upon GPA as a measure of transcript strength, particularly when students systematically select into courses based on ability. Through both empirical analysis and simulation evidence, we show the conditions under which IRT approaches outperform GPA, suggesting researchers should consider IRT-based alternatives when analyzing transcript data.

Specifically, our analysis yields several key findings. First, we show that IRT-estimated course difficulties align with conventional understanding, with AP/IB STEM courses being the most challenging and applied courses the least challenging. Second, we demonstrate that $\hat{\theta}$ captures meaningful variation in student performance beyond GPA, particularly for students who achieve the same GPA through different course-taking patterns—taking either more difficult or easier courses. Third, $\hat{\theta}$ is a stronger predictor of college outcomes than either GPA or SAT scores, maintaining its predictive power even after controlling for both measures. Fourth, through simulation studies, we show that when students take even a small common set of courses, IRT-estimated $\hat{\theta}$ better recovers true student ability than GPA in the presence of ability-based selection into courses. Having validated our methodology, we then examine how different measures characterize educational inequality. We find that compared to the SAT, $\hat{\theta}$ shows smaller racial/ethnic gaps but substantially larger gender gaps that indicate boys underperform in their courses relative to their standardized test scores.

Situating the Study

Our study addresses three related strands of literature. The first is the signaling value of a high school transcript in the context of SAT test-optional admissions policy. The field of educational measurement has recently grappled with the fairness and consequences of college admissions tests (Ackerman, 2021; Albano, 2021; Briggs, 2021; Franklin et al., 2021; Geisinger, 2021; Klugman et al., 2021; Koretz, 2021; Lyons et al., 2021; Mattern et al., 2021; McCall, 2021; Randall, 2021; Torres Irribarra & Santelices, 2021; Walker, 2021; Way & Shaw, 2021). If required, college admissions tests certainly play a role in the admissions process and can produce inequitable outcomes given unequal opportunities to develop skills (e.g., as exemplified by cross group differences in test scores) and test preparation (Briggs, 2021). Yet, counterfactually, much will depend on the information contained in other admissions materials if standardized test scores are omitted. Rothstein (2022) shows that recommendation letters differ in quality by racial and socioeconomic subgroups but do not contribute much weight to admissions decisions. What is likely but not definitively known is that the high school transcript is likely to have a larger influence on admissions decisions when test scores are omitted. Then, the question becomes: what latent information about student academic performance in high school is contained in a transcript? Our paper seeks to provide that answer.

The second related strand is measurement. To our knowledge, our study is the first to apply IRT to estimate the relative strength of each transcript in a full state-level dataset of high school transcripts and to demonstrate the construct and predictive validity of the measure, where "construct validity" is the extent to which the measurement tool (e.g., IRT-based transcript strength score in our case) represents the theoretical concept it is intended to describe (Shadish et al., 2002), and "predictive validity" is the extent to which the measure is associated with future outcomes as expected – what is often called "validity based on relations to other variables" in educational and psychological measurement (AERA et al., 2014)[1].

We identified three previous applications of IRT to transcript data. Most recently, Hansen et al. (2019) used IRT to establish the relative difficulty of a set of high school math courses, though their analysis was limited to a population of first-year calculus students, 20 courses in math and science only, self-report course grades, and minimal emphasis on validating the measure empirically or theoretically. Bassiri & Schulz (2003) used an IRT-based approach to establish a scale for high school course difficulty, though they intended to see whether transcript data could replicate ACT scores and therefore linked the ACT to their estimation procedure. This approach not only relies on the assumption that high school courses measure the same construct as the ACT (an assumption we disprove below) but also obviates the possibility of comparing what unique information a transcript provides relative to a standardized test. Similarly, Lei et al. (2001) used IRT to overcome the non-comparability of GPAs in different *college* courses, but for a limited sample of students at two universities and with limited effort to validate. Still, their findings inform our methods in that they suggest the use of the Partial Credit Model (PCM; Masters, 1982) over the more complex Graded Response Model (GRM; Samejima, 1969) and Generalized Partial Credit Model (GPCM; Muraki, 1992) due to model stability issues with the more complex models.

[1] We note here that while we draw upon the AERA, APA and NCME joint standards here, we intentionally do not use the standards' full framework for validity evidence, as the standards are written to describe evidence requirements for test *development*, whereas this study is a secondary data analysis.

The third related strand is the utility of GPA as a signal of student academic performance. As is well known, high school GPA plays a crucial role in several major educational policies around the world. In Chile, university admissions use a GPA rank score as part of their centralized admission system, with evidence suggesting this policy has led to strategic responses in high school grading practices (Fajnzylber et al., 2019). In the United States, the Texas Top 10% Rule guarantees state university admission to students graduating in the top decile of their high school class, a policy that has significantly impacted college enrollment patterns and serves as an alternative to race-based affirmative action (Cortes, 2010; Niu & Tienda, 2010). State merit aid programs also frequently use GPA requirements, with Georgia's HOPE scholarship program being a prominent example requiring students to maintain a 3.0 GPA. This program has had substantial effects on institutional behavior and student outcomes (Long, 2004).

Relatedly, the GPA is used widely in empirical research despite its self-evident limitations as a measure of student performance in high school. Researchers commonly employ GPA either as a covariate/predictor of other outcomes (Cohn et al., 2004; Grove et al., 2006) or as an outcome variable itself (Backes et al., 2024, 2024; Goldhaber & Goodman Young, 2024; Hill, 2015; Jackson, 2018). In general, GPA serves as a proxy for the latent variable of interest—student performance in high school, or what we refer to as transcript strength. For example, researchers use GPA to assess whether students have *learned* more because of a policy (GPA as outcome) or to control for selection into a program by adjusting for *student performance* (GPA as covariate). However, GPA likely suffers from correlated measurement errors, as lower-ability students may select into easier courses, inflating their GPA, while higher-ability students may choose harder courses, deflating their GPA. Economists have long recognized that using variables with non-classical measurement error on either side of regression equations can result in unpredictable bias (Pischke, 2007). Yet, perhaps due to the absence of available alternatives, researchers have used GPA as a proxy for transcript strength with little critical attention. In this paper, we aim to provide a feasible alternative to GPA using IRT methods, which may be of use to policymakers as well desiring to assign program opportunities based on high school performance and are worried that the GPA can mask differences in performance across course difficulty.

Data

Our analytic sample consists of 43,767 students from Delaware from graduating cohorts 2017 to 2021, with descriptive statistics shown in Table 1. The sample is diverse, with 48% White students, 30% Black students, 15% Hispanic students, and 4% Asian students. Approximately one-third (34%) of students are from low-income backgrounds, while 13% are students with disabilities (SWD) and 13% are English language learners (ELL). The gender distribution is nearly even, with 49% male students. Academic performance indicators show that students earned a mean GPA of 2.91 (SD=0.73) and achieved average SAT scores of 476.42 (SD=100.33) in Mathematics and 488.76 (SD=99.54) in English Language Arts, with the ELA sample slightly smaller at 43,744 students.

< Table 1 Here >

In Table 2, we show the top 5 percent of enrolled courses. No course is enrolled by all students, but some are enrolled by nearly everyone. For example, Health Education and Physical Education have 39,850 (91%) and 38,488 (88%) of all students, respectively. Core academic subjects also show high enrollment numbers, with Biology (37,717 students), English/Language Arts I (36,703 students), and English/Language Arts II (35,229 students) rounding out the top five most enrolled courses. Course performance, as measured by mean GPA, varies considerably among these subjects with high enrollment. Mean GPA is 3.3 in these courses with a standard deviation of 0.27.

Unsurprisingly, Physical Education, Career Exploration and Fitness/Conditioning have the highest GPAs, whereas core academic subjects have GPAs ranging from 2.5-2.8.

<Table 2 Here>

Methods

Overview of the Item Response Theory (IRT) Framework

The measurement approach used in this study is based upon the premise that when external readers such as admissions officers review applicants' high school transcripts, they make comparisons of the overall strength or quality of the transcripts according to a holistic judgment that considers the entirety of the transcript. In making such judgments, the reader must make comparisons between students who have taken different sets of courses in different schools, requiring some degree of informal equating of the difficulty of different courses in order to arrive at the final judgment. Given this framing, we estimate transcript strength as a continuous, unidimensional, latent variable using an IRT model that takes students' recorded grade in each course as input and produces transcript strength and course difficulty estimates on a common scale. In short, we treat each transcript like a set of answers to questions on a test and try to estimate the difficulty of each "item" so that we can ascribe proper value to an A grade in an easier course versus an A (or B, C, etc.) grade in a harder course for the purpose of summarizing the strength of each transcript as a score. Given the wealth of accessible guides to IRT in general and Rasch modeling specifically (e.g. Baker & Kim, 2004, 2017; Bandalos, 2018; Bond & Fox, 2015; Wilson, 2023), we focus here on what IRT has to offer in the analysis of transcript data.

IRT emerged from the field of educational testing (Birnbaum, 1968, in Lord & Novick, 1968; Lord, 1980) to address several issues inherent in the use of number- or proportion-correct scores on tests (this extends to scores based on average item scores, such as GPA). Many of the issues with proportion-correct scoring boil down to the fact that students taking two different tests do not inherently have comparable scores due to differences in the difficulty of the items on the two tests, even if the tests are built to assess very similar content. Most importantly for the present application, when items are calibrated as part of the same model, IRT produces item parameters and examinee scores on the same scale even if groups of students take different items. For example, a large-scale testing program might use one form in year 1, and a different form in year 2 for security. As long as the items were initially calibrated as part of the same model, no additional work would be needed to produce scores on a common scale, unlike in classical linking and equating contexts (Kolen & Brennan, 2014). This is possible because the item parameters in IRT are an integral part of scoring. Proportion-correct scores do not directly represent the difficulty of the items in any way, while IRT-based scores do.

The IRT framework, relative to GPA alone, provides two notable advantages. First, it produces estimates of the difficulty of items (in this case, courses) that are used to score students. A course might appear difficult in terms of GPA because the course is challenging, because the course is taken by generally lower-performing students, or both. IRT methods (under certain conditions, discussed below) can overcome this, while GPA cannot because GPA itself does not in any way account for the difficulty of the courses in which the grades being averaged were observed.[2] IRT's definition of a scale for the items then carries through to produce scores – estimates of the strength of an examinee's standing on a continuum representing the construct being measured by the test,

[2] This is not strictly true with GPA, as GPA is sometimes weighted to reflect greater course difficulty, such as A's in AP courses being worth 5 points towards GPA in some high schools. These weights however are arbitrarily selected for a handful of courses and discrete, not continuous, and therefore the general point about GPA not differentiating between course selection stands.

typically referenced as θ. Estimates of θ account for the difficulty of the items taken, a feature that makes IRT potentially more appropriate for summarizing transcripts than GPA.

Second, IRT can be used to produce reliable and unbiased estimates of course difficulty – the foundation for generating student scores – even if students have not answered the same questions (or, in our case, taken the same courses) (Thissen & Orlando, 2001). This is possible as long as respondents overlap in some of the items they answer, as demonstrated in the multistage adaptive testing literature, in which students are provided specific items calibrated to their performance on previous items in blocks (Glas, 1988; Steinfeld & Robitzsch, 2021; Wang et al., 2020), and explored in this study via simulation.

The IRT framework provides an estimate of college readiness based on students' high school transcript data. While this approach offers advantages over traditional GPA measures, it still faces certain limitations. An IRT approach cannot fully account for variations in course content, teaching practices, and school-specific grading policies such as grade inflation. However, these limitations are unlikely to be more severe than those inherent in using GPA alone. The fundamental challenge lies in synthesizing complex educational data: students take diverse courses across different schools with varying grading practices, and their course selection decisions depend heavily on specific contexts. The IRT methodology addresses this challenge by reducing this multifaceted data into a single continuous unidimensional variable. This approach mirrors, in a formal and standardized way, the process that college admissions officers already undertake informally when they evaluate transcripts to make admissions decisions. While individual institutions may codify this evaluation process differently, they all face the need to transform complex transcript data into actionable insights. IRT provides a systematic framework for this transformation.

Model specification.

To produce student-level transcript strength estimates and course-level difficulty estimates on a common scale, we calibrate the PCM. The PCM is a Rasch-type (Rasch, 1960) IRT model for polytomous data (i.e. categorical data with more than two ordered response categories). The PCM models the probability of responses from a person *n* to an item *i* in category *k* as a function of a person-side parameter θ_n representing, in this case, the overall strength of a student's high school academic performance, as well as the overall difficulty of an item δ_i and a threshold τ_{ik}. The item response function for the PCM is

$$\ln\left[\frac{P_{n_i}(xi = k)}{P_{n_i}(xi = k - 1)}\right] = \theta_n - \delta_i - \tau_{ik} \qquad (1)$$

That is, the log-odds of a response to item *i* in category *k*, relative to category *k* – 1, are a linear function of the three parameters specified above. The higher an individual's θ_n and/or the lower the overall difficulty of the item δ_i and threshold τ_{ik}, the likelier the individual is to respond in category *k* compared to the category one below *k*. That is, the PCM is a linear model in the logit metric.

In this parameterization, each τ_{ik} expresses the location of the threshold relative to δ_i. For example, say a given course has a δ_i of 1 and a τ_{i2} of 0.5, representing the threshold between category 1 and category 2. The point on the logit scale at which responses of 1 and 2 are equally likely is $\delta_i + \tau_{i2} = 1.5$. As such, the PCM is often reparametrized to remove the δ_i parameter entirely and just estimate threshold parameters. In this reparameterization, the example τ_{i2} would be 1.5, and no δ_i would be estimated. This is the parameterization we use in this study.

<Figure 1 here>

The parameters of the PCM combine to form item characteristic curves (ICCs) such as that shown in Figure 1. This figure contains four curves, each representing the probability of responding in a given category at the θ value on the x-axis. The points at which the curves representing probabilities for adjacent categories cross are the thresholds, such that for example the left-most crossing is the threshold between responding in category 1 and responding in category 2. At this value of θ, a bit below -2, a respondent would have an equal probability of responding in category 1 versus category 2. As θ increases, the probability of a response in category 1 decreases, while the probability of a response of 2 increases – to a point. The probability of a 2 starts decreasing as θ approaches the next threshold because as θ continues to increase, a response of 2 starts to become less likely and a response of 3 (or 4, the highest category) becomes increasingly likely.

Model estimation.

We estimate the PCM using full-information marginal maximum likelihood (FIMML) as implemented in the R software package *mirt* (Chalmers, 2012). The full-information aspect of estimation reflects the fact that the estimation procedure accounts for all individual response strings, a critical feature given the huge number of distinct course-taking patterns present in our data (in contrast to limited-information approaches based upon e.g. a variance-covariance matrix; see Cai, 2010 for an outline of the estimation algorithm implemented in *mirt*); the "marginal" aspect reflects the fact that estimation is based upon assuming a distribution for θ (in this and most cases, a normal distribution).

Though full-information maximum likelihood estimation typically assumes that data are missing at random or completely at random (Enders & Bandalos, 2001)—an assumption clearly unlikely to hold in a course-selection context where students choose which courses to take as a function of what their school offers, what interests them, and other unobserved selection mechanisms—the implications of this assumption for estimation of the PCM are not necessarily as problematic as one might suspect. Notably, IRT models estimated with FIMML have proven to be remarkably robust to nonrandom missingness of item responses. This robustness is most clearly demonstrated by the fact that one can readily estimate unbiased item parameters from a multistage adaptive test (Glas, 1988; Steinfeld & Robitzsch, 2021; Wang et al., 2020) using FIMML. This result occurs because the multistage design involves a subset of items taken by many or all examinees, enabling the estimation procedure to effectively anchor the item parameters against this "core" of courses for which examinees' performances can be compared directly. More detail on parameter estimation for IRT models is available in Baker & Kim (2017).

The PCM places individuals and items onto a scale that is common, but whose location on the continuum from negative to positive infinity is indeterminate. That is, in Rasch-type models, the distances between the item parameters and θ variance are identified from the item response data, but the mean of the θ and difficulty parameter distributions are not (Bechger & Maris, 2015). One must therefore fix a parameter in the model to "anchor" the scale and identify the model. We follow the *mirt* default (and common choice in IRT) of specifying 0 as the mean of the θ distribution. Importantly, a choice of a different anchor (e.g. specifying the mean item difficulty instead of the mean θ or choosing a different value for the mean of θ, such as 1000) would have no impact on model fit or any of our subsequent analyses, as all parameters would be shifted by the same linear transformation.

Meta-Analytic Average of Item Parameters

For each course *i*, there are typically 4 thresholds (j=1,2,3,4 for D,C,B,A) *j* associated with the course. To calculate average course difficulty, we take a meta-analysis approach by averaging threshold parameters for each item, weighting the average by the estimated variances of the thresholds. Let τ_{ij} represent the threshold parameter estimate from course i for threshold j, with associated standard error σ_{ij}. Then, the random effects meta-analysis framework takes the mean threshold parameter $\overline{\tau_{\imath\jmath}}$ for each course type i and threshold j by assuming that $\tau_{ij} = \overline{\tau_{\imath\jmath}} + \varepsilon_{ij}$, where ε_{ij}represents sampling error (in this case, the standard error of the parameter). Then, the meta-analytic estimate $\overline{\tau_{\imath\jmath}}$ is computed as $\overline{\tau_{\imath\jmath}} = \Sigma_j(w_{ij}\tau_{ij})/\Sigma_j w_{ij}$, where w_{ij} are inverse variance weights $w_{ij} = 1/\sigma_{ij}^2$.

Scoring.

Model estimation as outlined above produces item parameters, but not individual θ estimates (these are marginalized out of the estimation procedure, as indicated by the term “marginal” maximum likelihood). Producing θ estimates is therefore a separate step from estimating item parameters. Note that when we refer to θ estimates in the analyses that follow, we reference them as $\hat{\theta}$, in line with the fact that these scores are estimates of the unobservable underlying value θ for each student.

We compute $\hat{\theta}$ via maximum likelihood (MLE; see Thissen & Orlando, 2001). MLE scoring assigns to each student the value of $\hat{\theta}$ that maximizes the likelihood of their observed course grades in the courses they took, treating the estimated parameters of the PCM as known. MLE is often presented in contrast to empirical Bayes scoring methods such as expected a posteriori (EAP) scoring (Bock & Mislevy, 1982). EAP scoring assigns to each individual, as their estimated score, the mean of the posterior distribution produced by multiplying together the probability distributions representing (1) their individual $\hat{\theta}$ likelihood, given their item responses, and (2) the population-level first and second moments of a normal θ distribution. The main benefit of EAP scoring relative to maximum likelihood is that the use of a population prior distribution makes it possible to produce a $\hat{\theta}$ for individuals with all-perfect or all-zero response strings; maximum likelihood estimates for these patterns would be positive and negative infinity, respectively. However, for relatively short tests, EAP (and empirical Bayes scoring, more generally) can substantially shrink the distribution of $\hat{\theta}$s toward zero, the default mean of the θ distribution. During preliminary analysis of the models used in this study, we found that EAP scoring produced an unacceptable amount of shrinkage, especially for students with very high course grades. On this basis, we opted to use MLE scoring.

This means that a strategy for dealing with "perfect" (i.e. all-A) or "zero" (i.e. all F) course grades was required. We considered several strategies for addressing this issue. Arbitrary definition of a highest and lowest obtainable $\hat{\theta}$ was deemed unacceptable due to this approach failing to differentiate between perfect grades in easy versus difficult courses. Typical IRT methods based upon the test characteristic curve (Lord & Wingersky, 1984) were deemed infeasible because there are thousands of different ways that students' transcripts combine the 700+ courses in the dataset, and each observed combination with at least one perfect or all-zero transcript would need to be treated as its own test with its own characteristic curve. We ultimately settled upon a somewhat arbitrary approach of, for perfect course grades, recoding the student's grade of A to B for the hardest course they took (i.e. the course in their transcript with the highest τ_4); Linacre (2009) provides some precedent for this scoring approach. This has the effect of slightly shrinking $\hat{\theta}$ estimates for these students toward 0, but far less so than EAP scoring; the students with these course grades remained very high in the $\hat{\theta}$ distribution after recoding.

Course grades as item responses.

The large majority of course grades in the dataset follow an A-B-C-D-F scheme, and we limited our analytic dataset to grades following this scheme. We coded these grades numerically as A = 4, B = 3, C = 2, D =1, F = 0. For courses that students took more than once, we assigned the mean of their multiple numerically coded course grades, rounded to the nearest integer, as their course grade for the purpose of fitting the model and producing scores. For example, a student who took a course three times and received grades of D, A, A would have a score of 3 for that course in our analytic dataset.

We further limited courses in the analytic sample to those taken by at least 10 of these students, and in which at least two distinct grades were given according to our 0-3 grading scheme. These choices were made to ensure that thresholds could be estimated for all courses in the analytic dataset. We conducted extensive checks of the appropriateness and fit of the model; a summary of which is available in Appendix A: IRT Model Fit Results.

Results

Construct Validity: Plausible course difficulty estimation.

We now turn to estimation results. We start by looking at course difficulty, focusing on the average course difficulty $\overline{\tau_{ij}}$ instead of individual item parameters. Because we have prior understanding of "hard" and "easy" courses, the item parameters provide a useful piece of construct validation: if the IRT model orders course difficulty as we might expect, this lends credibility to the exercise. We start with Figure 2, which shows a meta-analytic ranking of course difficulties. The results largely align with conventional understanding of course difficulty hierarchies, providing important construct validation for our methodology. Advanced Placement (AP) courses, particularly in STEM subjects, cluster at the higher end of the difficulty spectrum, with AP Physics C emerging as the most challenging course (≈+4 logits). This is followed by other AP science and mathematics courses. At the other end of the spectrum, more applied and introductory courses such as Workplace Experience, Physical Education, and basic ROTC courses show lower difficulty estimates (≈-4 logits). Core academic subjects generally fall in the middle range, with some variation by level and content area.

The color-coding of courses by subject area reveals interesting patterns in the difficulty hierarchy. STEM courses (shown in blue shades) tend to concentrate in the upper half of the difficulty range, while arts and sports (orange) typically appear in the lower half. Language courses (maroon) show a clear progression of difficulty, with advanced language courses and AP language offerings positioned higher on the scale than their introductory counterparts. This systematic ordering of courses by difficulty level, which aligns with educational experience and prior research, suggests that our IRT-based methodology effectively captures meaningful differences in course challenge levels.

<Figure 2 Here>

Looking across all courses confirms these patterns. To do this, we apply a meta-analytic regression to the average course difficulties leveraging the fact that for each course difficulty we estimate its cross-item parameter variance. We then use binary predictors to explain variance in course difficulty across subject areas and course levels (e.g., AP/IB versus regular courses, indicator variable for course content areas). The regression results reveal several key patterns. First, compared to mathematics courses (the reference category), arts and sports courses are significantly less difficult, with coefficients (in logits) of -1.005 and -1.021 respectively ($p < 0.001$). Science, humanities, and language courses do not differ significantly from mathematics in their difficulty levels, with relatively small and statistically insignificant coefficients ranging from -0.107 to 0.078.

Advanced Placement (AP) and International Baccalaureate (IB) courses consistently show higher difficulty levels across all subject areas. The AP premium is largest in mathematics (1.748) and science (1.588), followed by humanities (1.297), arts (0.973), and languages (0.829). Similarly, IB courses show substantial increases in difficulty across most subjects, with effects ranging from 1.410 in mathematics to 1.101 in languages, though notably, the IB effect in humanities (0.246) is not statistically significant. These findings quantify the considerable increase in academic challenge represented by AP and IB coursework, while also highlighting how this challenge premium varies across subject areas. More broadly, these results show that IRT-estimated item difficulties match common understanding of course difficulty.

<Table 3 Here>

Construct Validity: Plausible transcript strength ($\hat{\theta}$) estimation.

Having established the validity of our course difficulty estimates, we now examine student transcript strength ($\hat{\theta}$) and its relationship to GPA. While we expect $\hat{\theta}$ and GPA to be highly correlated, $\hat{\theta}$ should provide additional information particularly when students select into more challenging courses. Figure 3 illustrates this relationship through three complementary analyses.

Panel A shows the variation in $\hat{\theta}$ for students with GPAs of 2.0, 2.5, 3.0, and 3.5. For each GPA level, we observe substantial variation in transcript strength ($\hat{\theta}$), which we divide into terciles for subsequent analysis. This variation suggests that students achieving the same GPA do differ meaningfully in their underlying academic performance. Panel B demonstrates how this performance difference manifests in course-taking patterns by plotting the difference in enrollment rates between top (Tercile 3) and bottom (Tercile 1) $\hat{\theta}$ students across courses of varying difficulty. The consistently positive slopes across all GPA bands indicate that higher-ability students systematically select into more challenging courses, regardless of their GPA level.

Panel C reveals the consequences of this differential course-taking behavior. We plot the cumulative contribution to GPA as students progress through increasingly difficult courses, separated by $\hat{\theta}$ tercile within each GPA band. While higher-$\hat{\theta}$ (Tercile 3) students initially maintain higher GPAs, their tendency to enroll in more challenging courses gradually reduces their GPA advantage, ultimately converging with their lower-$\hat{\theta}$peers who take easier courses. This pattern explains the substantial within-GPA variation in $\hat{\theta}$ observed in Panel A and demonstrates how IRT-based ability estimates capture important information about student achievement that GPA alone masks. Specifically, students with the same GPA may arrive at that outcome through different pathways: higher-ability students taking more challenging courses versus lower-ability students taking less challenging ones.

<Figure 3 Here>

Construct Validity: $\hat{\theta}$ and SAT comparisons.

Having now shown that item difficulties capture meaningful representations of course difficulty and student ability ($\hat{\theta}$) captures meaningful representations of student selection into more or less difficult courses, we now turn to substantive analysis of $\hat{\theta}$ and compare it to the other primary metric college admissions officers use to evaluate student ability, the SAT. Figure 4 illustrates the relationship between students' standardized $\hat{\theta}$ scores and their average Math and ELA SAT scores. While there is a clear positive relationship between these measures (adjusted $R^2 = 0.51$, implied correlation of 0.71), there remains substantial unexplained variation. A 100-point increase in SAT score is associated with a 1.25 standard deviation increase in $\hat{\theta}$, suggesting these measures capture related but distinct aspects of student achievement. The marginal distributions shown along the axes reveal that both measures are approximately normally distributed, though $\hat{\theta}$ shows somewhat heavier tails, particularly in identifying high-achieving students.

<Figure 4 Here>

As with the $\hat{\theta}$/GPA comparison, residual variation in $\hat{\theta}$ conditional on the SAT (and vice-versa) can be used to understand how a student's transcript can contain ability-related information distinct from the SAT itself. To do this, we divide students into four groups based on their joint SAT and $\hat{\theta}$ distributions: high-SAT/high-$\hat{\theta}$, high-SAT/low-$\hat{\theta}$, low-SAT/high-$\hat{\theta}$, and low-SAT/low-$\hat{\theta}$, where "high" and "low" are defined by the top and bottom two quintiles respectively.

Results shown in Table 4 reveal striking patterns in both course-taking behavior and performance across these groups. High-SAT/high-$\hat{\theta}$ students (N=12,477) show consistently strong performance across all courses (GPAs typically above 3.5) and high enrollment rates in advanced courses. For instance, 55% take Pre-Calculus and 24% take AP Calculus AB, with GPAs of 3.37 and 3.33 respectively. In contrast, high-SAT/low-θ students (N=2,067) show markedly lower GPAs despite similar SAT scores. While they maintain relatively high enrollment in standard courses, their participation in advanced courses also drops substantially (36% in Pre-Calculus, 4% in AP Calculus AB), and their GPAs in these courses are notably lower (1.79 and 1.35 respectively).

Perhaps most revealing are the low-SAT/high-$\hat{\theta}$ students (N=1,867), who demonstrate strong academic performance despite lower standardized test scores. These students achieve GPAs comparable to their high-SAT peers in many courses (e.g., 3.18 in AP Chemistry compared to 3.51 for high-SAT/high-θ students), though they enroll in advanced courses at lower rates. The low-SAT/low-θ group (N=12,302) shows both the lowest enrollment rates in advanced courses and the lowest GPAs across nearly all courses, suggesting that both measures capture meaningful dimensions of academic preparation.

These patterns suggest that $\hat{\theta}$ captures important information about student ability that complements SAT scores, particularly in identifying students whose classroom performance may diverge from their standardized test performance.

<Table 4 Here>

Predictive Validity: GPA, SAT, and $\hat{\theta}$ as predictors of college outcomes.

We obtain data from the National Student Clearinghouse (NSC) to investigate whether a student's GPA, SAT score, or $\hat{\theta}$ differently predict college-going behavior. We examine three college outcomes: expected earnings based on college attendance using the College Mobility Report Card (Chetty et al., 2017), college selectivity (coded 1-5, from no college to most selective institutions), and college completion (coded 1-4, from no college to 4-year degree), which is restricted to the 2017 graduating cohort.

Results shown in Table 5 reveal several key patterns. In Panel A, where GPA and $\hat{\theta}$ are estimated separately, both measures significantly predict college outcomes, but with different magnitudes. A one standard deviation increase in $\hat{\theta}$ is associated with a \$7,148 increase in expected earnings, compared to \$6,291 for GPA and \$5,132-\$6,128 for SAT (depending on model specification). For college selectivity and degree completion, $\hat{\theta}$ also shows stronger predictive power than both GPA and SAT, with odds ratios of 2.50 and 3.21 respectively, compared to 2.37 and 2.85 for GPA and lower values for SAT. All differences between coefficients are statistically significant ($p < 0.001$).

Panels B and C examine $\hat{\theta}$'s predictive power conditional on GPA, using two different approaches to GPA fixed effects. Panel B uses fixed effects for specific GPA values (2.0, 2.5, 3.0, 3.5), mirroring Figure 2 above, while Panel C uses a more granular approach with fixed effects for each rounded GPA value from 1.0 to 4.0 in increments of 0.1. In both specifications, $\hat{\theta}$ remains a strong predictor of college outcomes even after controlling for GPA. The effects are particularly striking for expected earnings, where a one standard deviation increase in $\hat{\theta}$ is associated with \$25,235-\$28,980 higher expected earnings, substantially larger than the SAT's effect of \$3,566-\$4,209. Similar patterns emerge for college selectivity and degree completion, with θ showing significantly stronger predictive power than SAT across all specifications ($p < 0.001$).

These results suggest that $\hat{\theta}$ captures important dimensions of academic preparation and potential that are distinct from both GPA and SAT scores, and these differences have meaningful implications for college outcomes. In Appendix Table B1, we take the groups from Table 4 – e.g., high SAT/low $\hat{\theta}$—and estimate whether those students with high residual $\hat{\theta}$ scores have different college outcomes than similarly scoring SAT students with low residual $\hat{\theta}$. As expected, high $\hat{\theta}$ students attend more selective colleges than low $\hat{\theta}$ students with similar SAT scores. Most notably, high $\hat{\theta}$-low SAT students attend more selective colleges than students in the low $\hat{\theta}$-high SAT students, furthering evidence that $\hat{\theta}$ is a better predictor of college outcomes than the SAT.

<Table 5>

Robustness to Grade Inflation

In Appendix C: Grade Inflation, we evaluate the influence of systematic grade inflation on $\hat{\theta}$. We note at the onset that a measure of transcript strength should be sensitive to grade inflation, a phenomenon that generally reduces the utility of transcript data (Tyner and Gershenson, 2022). Here, we focus on the extent to which grade inflation occurs mostly between schools or course types and whether $\hat{\theta}$ is more explained by grade inflation or an independent measure of ability, the SAT. First, stylistically, we show that there are schools that grade calculus courses differently, ranging about 0.5 GPA points on average, conditional on mathematics SAT. School-level $\hat{\theta}$ scores are similarly variable conditional on mathematics SAT, suggesting that these grading practices influence $\hat{\theta}$ estimation (Figure C1). Then, we generate a grade inflation variable and show that, on average, grade inflation predicts $\hat{\theta}$, but student mean SAT score is a better predictor. In standard deviation units, a 1 SD increase in student level grade inflation is associated with a 0.36 SD increase in $\hat{\theta}$, whereas a 1 SD increase in mean SAT is associated with a 0.72 SD increase in $\hat{\theta}$ (Table C1). We also cannot rule out other explanations; for example, residual GPA performance in a course may not be grade inflation but better teaching or greater student effort. Thus, though we conclude that grade inflation is likely to have some influence on latent transcript strength, an independent measure of student college readiness via the SAT more closely tracks the quality of a student's coursetaking history.

Course Selection Effects: Simulations of Course-Selection Bias for GPA and $\hat{\theta}$

To evaluate how student course selection patterns might affect our θ estimates, we conduct three simulation studies varying in their data generating processes (see the Appendix D: Simulation Study Details for more information). Each study simulates θ for 1000 students and course-taking behavior across 200 potential courses, but we estimate $\hat{\theta}$ using only 28 selected courses to mirror real-world scenarios where students take a subset of available courses. The studies vary in the distribution of course difficulty relative to θ and the mechanism by which courses are selected at the student level. In all cases, the average course difficulty is lower than the average θ, reflecting the distributions found when we fit the model to real data. For each study, we compare the extent to which $\hat{\theta}$ based on the 28 non-randomly selected courses recovers $\hat{\theta}$ based on all 200 courses, to the extent to which GPA from the 28 courses recovers GPA from all 200. We also report correlations between the "observed" $\hat{\theta}$ and GPA, and the extent to which the standardized "full" (all 200 courses) $\hat{\theta}$ and GPA recover the standardized data-generating θ.

Study 1 represents the simplest case, where course selection depends on $\theta$and courses vary in difficulty such that all students can take courses whose difficulty is close to their θ. Importantly, this scenario includes a "core" of 8 courses from which students must take at least five. Study 2 introduces range restriction in course difficulty while maintaining the core course requirement. Study 3 presents the most challenging scenario, combining restricted course difficulty range with no core course requirement. For each study, we conducted 50 replications.

Results shown in Table 6 demonstrate that IRT-estimated $\hat{\theta}$ generally outperforms GPA across all scenarios. In all studies, $\hat{\theta}$ estimates show minimal bias (ranging from -0.006 to 0.027) compared to larger biases in GPA (ranging from 0.104 to 0.181). The concordance correlation coefficients (Lin, 1989) for $\hat{\theta}$ are substantially higher than for GPA in Study 1 (0.967 vs 0.456) and Study 2 (0.979 vs 0.815). Study 3 demonstrates that there are circumstances under which IRT performs quite poorly (i.e. a relatively narrow range of course difficulty with no core of highly-taken courses), but GPA performs even worse – the concordance correlation coefficients are identical in this scenario, but the root-mean-square error of the GPA estimates relative to the GPAs observed when all 200 courses are observed is larger than the equivalent error for θ once the errors are normalized by the two measures' respective standard deviations.

Figure 5 provides visual confirmation of these patterns, plotting both $\hat{\theta}$ and GPA estimates from the 28-course subset against "true" values from all 200 courses. The bottom row shows that $\hat{\theta}$ estimates (blue curve) more closely follow the identity line (red) across all three studies, indicating strong recovery of true ability. In contrast, the top row shows GPA estimates deviate more substantially from the identity line, particularly at the extremes. This pattern is most pronounced in Study 3, where the absence of core courses leads to greater deviation in both measures, though $\hat{\theta}$ maintains better performance than GPA.

These results suggest that IRT-based ability estimation is sufficiently robust to various course selection patterns, particularly when students share some common courses, to be preferable to GPA in many cases. Even in more complex scenarios with restricted course difficulty and no core requirements, $\hat{\theta}$ provides more accurate transcript strength estimates than GPA alone. That is, even in data-generating scenarios that would be expected to make it challenging to do so, the IRT approach is able to account for course difficulty well enough to improve upon GPA (as GPA does not account for course difficulty in any way).

<Figure 5 Here>

<Table 6 Here>

$\hat{\theta}$ and SAT group-level inequality.

We now examine how $\hat{\theta}$ and SAT differently (or similarly) characterize between-group inequality. Since both $\hat{\theta}$ and SAT represent latent variables with indeterminate interval properties, we calculate ordinal achievement gaps using the V-Gap statistic of Ho (2009) and Ho and Reardon (2012), which can be interpreted as a standardized mean difference between groups under the assumption of respective normality. These gaps are visualized in Figure 6 through cumulative probability plots, where the area under the curve (AUC) represents the probability that a randomly selected student from one group scores higher than a randomly selected student from the comparison group. Further, we conduct tests to see whether the AUCs for $\hat{\theta}$ and SAT are different from each other.

The results reveal several distinct patterns across demographic groups. Most notably, male and female students have very different levels of inequality using the $\hat{\theta}$ metric, whereas for SAT scores, there is almost no difference ($p<0.000$). For the SAT, the AUC falls just below 0.5, meaning that males are just as likely to have higher scores as females. In contrast, the AUC in $\hat{\theta}$ is 0.377, meaning that a randomly drawn male student has just a 38% chance of having a higher score than a randomly drawn female student, corresponding to a Cohen's D-like difference (i.e., V) of 0.445. In contrast, socioeconomic gaps show remarkable consistency across both measures (p = 0.972), with nearly identical AUCs for non-low-income versus low-income students.

Racial/ethnic achievement gaps show varying patterns. The Asian-White gap is similar across both measures (p = 0.228), though with slightly different shapes in the probability curves. However, both Black-White and Hispanic-White gaps show significant differences between $\hat{\theta}$ and SAT measurements (p < 0.000 in both cases). For Hispanic students, the SAT shows larger gaps (V = -0.745) compared to θ (V = -0.386), while both measures indicate substantial gaps for Black students relative to White students, though with different magnitudes and distributions.

These findings suggest that choice of metric ($\hat{\theta}$ versus SAT) can substantially affect our understanding of educational inequities, particularly for gender and certain racial/ethnic comparisons. This has important implications for how we measure and interpret group-level differences in outcomes in educational policy and research.

<Figure 6 Here>

Standardizing $\hat{\theta}$ and the SAT allows us to compare estimates in a regression framework, though still reliant on the interval properties of the two scales. If, however, the standardized metric gaps are roughly commensurate to the V-gaps, which only rely on ordinal information and the assumption of respective normality, then we can be more confident that the interval properties of the respective metrics do not confound gap estimation. Standardizing then lets us test whether group-level differences in $\hat{\theta}$ remain controlling for SAT scores. This latter question is important as it speaks to whether $\hat{\theta}$ inequality is distinct from SAT inequality. We estimate group-level differences sequentially to avoid controlling for other group-level differences (e.g., estimating racial/ethnic differences controlling for economic differences). Results shown in Table 7 reveal several key patterns.

First, the standardized differences largely align with the ordinal gaps shown in Figure 6. For instance, gender differences are more pronounced in $\hat{\theta}$ (-0.413 SD) than in SAT (-0.062 SD). Unsurprisingly, since SAT scores do not appear to be correlated with gender, these differences persist even after controlling for SAT scores (-0.369 SD), suggesting that male and female students with equivalent SAT scores show substantially different course-taking and performance patterns.

Racial/ethnic differences show complex patterns. While Asian students show similar advantages over White students in both $\hat{\theta}$ (0.678 SD) and SAT (0.687 SD), a significant difference remains after controlling for SAT (0.202 SD). Black and Hispanic students show larger disadvantages in SAT (-0.763 SD and -0.651 SD respectively) than in $\hat{\theta}$ (-0.616 SD and -0.440 SD). However, after controlling for SAT scores, the Black-White gap substantially narrows (-0.088 SD) and the Hispanic-White gap becomes statistically indistinguishable from zero (0.010 SD).

Other notable findings include similar socioeconomic gaps across measures (-0.618 SD in $\hat{\theta}$, -0.633 SD in SAT) with a persistent gap even after controlling for SAT (-0.183 SD). Students with disabilities show larger gaps in SAT (-1.039 SD) compared to $\hat{\theta}$ (-0.748 SD), with no significant difference after controlling for SAT. English language learners show smaller gaps in $\hat{\theta}$ (-0.158 SD) compared to SAT (-0.389 SD) and, notably, show positive differences after controlling for SAT (0.121 SD), suggesting stronger course performance than their SAT scores would predict.

<Table 7 Here>

Discussion

Our findings suggest several promising directions for future research to enhance and validate transcript-based measures of student achievement. If estimation challenges can be overcome (Wang et al., 2020), random-effects differential item functioning models (Muthén & Asparouhov, 2018; Shear, 2018) could illuminate how course difficulty varies across educational contexts, while multidimensional IRT approaches (Reckase, 2009) might better capture domain-specific strengths. One could also explore the extent to which selection models (Du et al., 2022) might be applied to both further mitigate selection effects and model them as a formal phenomenon that can be studied unto itself. Perhaps most critically, research examining how $\hat{\theta}$ aligns with admissions officers' holistic transcript evaluations could validate whether our quantitative measure captures the latent construct that experienced evaluators assess qualitatively. Such studies could employ forced-choice comparisons or conjoint analyses to identify where algorithmic and human assessments diverge, particularly for historically marginalized groups.

The policy implications of our findings warrant careful consideration. While $\hat{\theta}$ demonstrates superior predictive validity compared to GPA, its potential application in high-stakes decisions like college admissions, or assignment to program benefits, as in the Texas Top 10 Plan, requires extreme caution. Campbell's law suggests that formalizing $\hat{\theta}$ as an admissions criterion could distort the very behaviors it aims to measure; for example, it may exaggerate already prevalent grade inflation patterns (Goldhaber and Goodman Young, 2024; Tyner and Gershenson, 2020). Additionally, differential access to advanced coursework means $\hat{\theta}$ may reflect opportunity as much as achievement, though this concern applies to standardized assessments also and may be more policy malleable. Nevertheless, our results indicate that transcript data contain valuable information about student preparation beyond what standardized tests capture, suggesting careful consideration of how to incorporate this information while mitigating potential adverse effects.

Several important limitations constrain interpretation of our results. First, $\hat{\theta}$ measures transcript strength rather than underlying ability - it cannot disentangle student capability from course availability or preference-based selection. Second, while GPA's interval-scale properties are highly questionable (Bond and Lang, 2013; Domingue, 2014), $\hat{\theta}$'s interval scaling depends on model fit, an area requiring further psychometric investigation. Third, non-random grade inflation may bias $\hat{\theta}$ estimates in ways that are difficult to anticipate or correct, though this phenomenon, too, is baked into transcript data and adjudication relies on admissions officers or codified institutional processes, a process that may be less credible or more biased than the formalized approach used here. These limitations underscore the importance of triangulating evidence across multiple measures and employing nonparametric analyses where possible.

Conclusion

This paper demonstrates that IRT methods can extract more information from high school transcripts than traditional GPA calculations, particularly when students systematically select courses based on ability. Our measure shows strong predictive validity for college outcomes while revealing different patterns of educational inequality than standardized tests. However, significant methodological and practical challenges remain before such measures could be considered for high-stakes applications. Future work should focus on understanding how transcript-based measures relate to expert judgment, vary across contexts, and might be made robust to strategic behavior. While our findings suggest promising directions for both research and practice, they also highlight the complexity of measuring and comparing student achievement across diverse educational settings.

References

Ackerman, P. L. (2021). Commentary: The Future of College Admissions Tests. *Educational Measurement: Issues and Practice*, *40*(4), 38–40. https://doi.org/10.1111/emip.12456

AERA, APA, & NCME. (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association.

Albano, A. D. (2021). Commentary: Social Responsibility in College Admissions Requires a Reimagining of Standardized Testing. *Educational Measurement: Issues and Practice*, *40*(4), 49–52. https://doi.org/10.1111/emip.12451

Backes, B., Cowan, J., Goldhaber, D., & Theobald, R. (2024). How to measure a teacher: The influence of test and nontest value-added on long-run student outcomes. *Journal of Human Resources*, 1023-13180R2. https://doi.org/10.3368/jhr.1023-13180R2

Baker, F. B., & Kim, S.-H. (Eds.). (2004). *Item Response Theory: Parameter Estimation Techniques, Second Edition* (Second edition). CRC Press.

Baker, F. B., & Kim, S.-H. (2017). *The Basics of Item Response Theory Using R*. Springer International Publishing. https://doi.org/10.1007/978-3-319-54205-8

Bandalos, D. L. (2018). *Measurement Theory and Applications for the Social Sciences*. Guilford Press.

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

Bechger, T. M., & Maris, G. (2015). A statistical test for differential item pair functioning. *Psychometrika*, *80*(2), 317–340. https://doi.org/10.1007/s11336-014-9408-y

Belasco, A. S., Rosinger, K. O., & Hearn, J. C. (2015). The test-optional movement at America's selective liberal arts colleges: A boon for equity or something else? *Educational Evaluation and Policy Analysis, 37*(2), 206-223. https://doi.org/10.3102/0162373714537350

Bennett, C. T. (2022). Untested admissions: Examining changes in application behaviors and student demographics under test-optional policies. *American Educational Research Journal, 59*(1), 180-216. https://doi.org/10.3102/00028312211003526

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Addison-Wesley.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*(4), 431–444. https://doi.org/10.1177/014662168200600405

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (3rd ed.). Routledge.

Bond, T. N., & Lang, K. (2013). The evolution of the Black-White test score gap in Grades K–3: The fragility of results. *Review of Economics and Statistics*, *95*(5), 1468-1479.

Briggs, D. C. (2021). Commentary: Comment on College Admissions Tests and Social Responsibility. *Educational Measurement: Issues and Practice*, *40*(4), 44–48. https://doi.org/10.1111/emip.12455

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika*, *75*(4), 581–612. https://doi.org/10.1007/s11336-010-9178-0

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chetty, R., Friedman, J. N., Saez, E., Turner, N., & Yagan, D. (2017). *Mobility report cards: The role of colleges in intergenerational mobility* (No. w23618). National Bureau of Economic Research.

Cohn, E., Cohn, S., Balch, D. C., & Bradley, J. (2004). Determinants of undergraduate GPAs: SAT scores, high-school GPA and high-school rank. *Economics of Education Review*, *23*(6), 577–586. https://doi.org/10.1016/j.econedurev.2004.01.001

Cortes, K. E. (2010). Do bans on affirmative action hurt minority students? Evidence from the Texas Top 10% Plan. *Economics of Education Review*, *29*(6), 1110–1124. https://doi.org/10.1016/j.econedurev.2010.06.004

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 837-845.

Dessein, W., Wouter, Frankel, A. & Kartik, N. The test-optional puzzle. (2025). *AEA Papers and Proceedings*. https://www.columbia.edu/~nk2339/Papers/DFK-TestOptionalPuzzle.pdf

Domingue, B. (2014). Evaluating the equal-interval hypothesis with test score scales. *Psychometrika*, *79*(1), 1-19. https://doi.org/10.1007/s11336-013-9342-4

Du, H., Enders, C., Keller, B. T., Bradbury, T. N., & Karney, B. R. (2022). A Bayesian latent variable selection model for nonignorable missingness. *Multivariate Behavioral Research*, *57*(2–3), 478–512. https://doi.org/10.1080/00273171.2021.1874259

Enders, C., & Bandalos, D. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*(3), 430–457. https://doi.org/10.1207/S15328007SEM0803_5

Fajnzylber, E., Lara, B., & León, T. (2019). Increased learning or GPA inflation? Evidence from GPA-based university admission in Chile. *Economics of Education Review*, *72*, 147–165. https://doi.org/10.1016/j.econedurev.2019.05.009

Franklin, D. W., Bryer, J., Andrade, H. L., & Lui, A. M. (2021). Commentary: Design Tests with a Learning Purpose. *Educational Measurement: Issues and Practice*, *40*(4), 64–65. https://doi.org/10.1111/emip.12457

Geisinger, K. F. (2021). Commentary: Social Responsibility, Fairness, and College Admissions Tests. *Educational Measurement: Issues and Practice*, *40*(4), 57–60. https://doi.org/10.1111/emip.12450

Glas, C. A. W. (1988). The Rasch model and multistage testing. *Journal of Educational Statistics*, *13*(1), 45–52. https://doi.org/10.3102/10769986013001045

Goldhaber, D., & Goodman Young, M. (2024). Course grades as a signal of student achievement: Evidence of grade inflation before and after COVID-19. *Journal of Policy Analysis and Management*, *43*(4), 1270–1282. https://doi.org/10.1002/pam.22618

Grove, W. A., Wasserman, T., & Grodner, A. (2006). Choosing a proxy for academic aptitude. *The Journal of Economic Education*, *37*(2), 131–147. https://doi.org/10.3200/JECE.37.2.131-147

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory*. Springer Netherlands. https://doi.org/10.1007/978-94-017-1988-9

Hansen, J., Sadler, P., & Sonnert, G. (2019). Estimating high school GPA weighting parameters with a graded response model. *Educational Measurement: Issues and Practice*, *38*(1), 16–24. https://doi.org/10.1111/emip.12203

Hill, A. J. (2015). The girl next door: The effect of opposite gender friends on high school achievement. *American Economic Journal: Applied Economics*, *7*(3), 147–177. https://doi.org/10.1257/app.20140030

Ho, A. D. (2009). A nonparametric framework for comparing trends and gaps across tests. *Journal of Educational and Behavioral Statistics*, *34*(2), 201–228. https://doi.org/10.3102/1076998609332755

Ho, A. D., & Reardon, S. F. (2012). Estimating achievement gaps from test scores reported in ordinal "proficiency" categories. *Journal of Educational and Behavioral Statistics*, *37*(4), 489–517. https://doi.org/10.3102/1076998611411918

Jackson, C. K. (2018). What do test scores miss? The importance of teacher effects on non–test score outcomes. *Journal of Political Economy*, *126*(5), 2072–2107. https://doi.org/10.1086/699018

Klugman, E. M., An, L., Himmelsbach, Z., Litschwartz, S. L., & Nicola, T. P. (2021). Commentary: The Questions We *Should* Be Asking About Socially Responsible College Admission Testing. *Educational Measurement: Issues and Practice*, *40*(4), 28–31. https://doi.org/10.1111/emip.12449

Kolen, M. J., & Brennan, R. L. (2014). *Test Equating, Scaling, and Linking* (3rd ed.). Springer New York. https://doi.org/10.1007/978-1-4939-0317-7

Koretz, D. (2021). Commentary: Response to Koljatic et al.: Neither a Persuasive Critique of Admissions Testing Nor Practical Suggestions for Improvement. *Educational Measurement: Issues and Practice*, *40*(4), 35–37. https://doi.org/10.1111/emip.12454

Lei, P.-W., Bassiri, D., & Schultz, E. M. (2001). *Alternatives to the grade point average as a measure of academic achievement in college*. American College Testing Program. https://files.eric.ed.gov/fulltext/ED462407.pdf

Lin, L. I.-K. (1989). A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics*, *45*(1), 255. https://doi.org/10.2307/2532051

Linacre, J. M. (2009). The efficacy of Warm’s weighted mean likelihood estimate (WLE) correction to maximum likelihood estimate (MLE) nias. *Rasch Measurement Transactions*, *23*(1), 1188–1189.

Long, B. T. (2004). How do financial aid policies affect colleges?: The institutional impact of the Georgia HOPE Scholarship. *Journal of Human Resources*, *XXXIX*(4), 1045–1066. https://doi.org/10.3368/jhr.XXXIX.4.1045

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. L. Erlbaum Associates.

Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Addison-Wesley.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score “equatings.” *Applied Psychological Measurement*, *8*(4), 453–461. https://doi.org/10.1177/014662168400800409

Lyons, S., Hinds, F., & Poggio, J. (2021). Commentary: Evolution of Equity Perspectives on Higher Education Admissions Testing: A Call for Increased Critical Consciousness.

Educational Measurement: Issues and Practice, *40*(4), 41–43. https://doi.org/10.1111/emip.12458

Mattern, K., Cruce, T., Henderson, D., Gridiron, T., Casillas, A., & Taylor, M. (2021). Commentary: Reviving the Messenger: A Response to Koljatic et al. (2021). *Educational Measurement: Issues and Practice*, *40*(4), 53–56. https://doi.org/10.1111/emip.12459

McCall, M. (2021). Commentary: Restoring Public Trust. *Educational Measurement: Issues and Practice*, *40*(4), 70–72. https://doi.org/10.1111/emip.12466

Muthén, B., & Asparouhov, T. (2018). Recent methods for the study of measurement invariance with many groups: Alignment and random effects. *Sociological Methods & Research*, *47*(4), 637–664. https://doi.org/10.1177/0049124117701488

Newman, D. A., Tang, C., Song, Q. C., & Wee, S. (2022). Dropping the GRE, keeping the GRE, or GRE-optional admissions? Considering tradeoffs and fairness. *International Journal of Testing*, *22*(1), 43–71. https://doi.org/10.1080/15305058.2021.2019750

Niu, S. X., & Tienda, M. (2010). The impact of the Texas top ten percent law on college enrollment: A regression discontinuity approach. *Journal of Policy Analysis and Management*, *29*(1), 84–110. https://doi.org/10.1002/pam.20480

Pischke, S. (2007). *Lecture notes on measurement error*. London School of Economics. https://econ.lse.ac.uk/staff/spischke/ec524/Merr_new.pdf

Randall, J. (2021). Commentary: From Construct to Consequences: Extending the Notion of Social Responsibility. *Educational Measurement: Issues and Practice*, *40*(4), 32–34. https://doi.org/10.1111/emip.12452

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Danish Institute for Educational Research.

Reckase, M. D. (2009). *Multidimensional Item Response Theory*. Springer New York. https://doi.org/10.1007/978-0-387-89976-3

Rothstein, J. (2022). Qualitative information in undergraduate admissions: A pilot study of letters of recommendation. *Economics of Education Review*, *89*, 102285. https://doi.org/10.1016/j.econedurev.2022.102285

Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Houghton-Mifflin.

Shear, B. R. (2018). Using hierarchical logistic regression to study DIF and DIF variance in multilevel data. *Journal of Educational Measurement*, *55*(4), 513–542. https://doi.org/10.1111/jedm.12190

Steinfeld, J., & Robitzsch, A. (2021). Item parameter estimation in multistage designs: A comparison of different estimation approaches for the Rasch model. *Psych*, *3*(3), 279–307. https://doi.org/10.3390/psych3030022

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test Scoring* (pp. 73–140). Lawrence Erlbaum Associates Publishers.

Torres Irribarra, D., & Santelices, M. V. (2021). Commentary: Large-Scale Assessment and Legitimacy beyond the Corporate Responsibility Model. *Educational Measurement: Issues and Practice*, *40*(4), 61–63. https://doi.org/10.1111/emip.12460

Walker, M. E. (2021). Commentary: Achieving Educational Equity Requires a Communal Effort. *Educational Measurement: Issues and Practice*, *40*(4), 73–75. https://doi.org/10.1111/emip.12465

Wang, C., Chen, P., & Jiang, S. (2020). Item calibration methods with multiple subscale multi-stage testing. *Journal of Educational Measurement*, *57*(1), 3–28. https://doi.org/10.1111/jedm.12241

Way, W. D., & Shaw, E. J. (2021). Commentary: An Evidence-Based Response to "College Admission Tests and Social Responsibility." *Educational Measurement: Issues and Practice*, *40*(4), 66–69. https://doi.org/10.1111/emip.12467

Wilson, M. (2023). *Constructing measures: An item response modeling approach* (2nd ed.). Routledge.

Tables

Table 1. Descriptive Statistics Analytic Sample

	Mean	SD
Asian	0.04	0.20
Black	0.30	0.46
Hispanic	0.15	0.36
White	0.48	0.50
Low-Income	0.34	0.47
SWD	0.13	0.33
ELL	0.13	0.34
Male	0.49	0.50
SAT - Math	476.42	100.33
SAT – ELA*	488.76	99.54
GPA	2.91	0.73

N=43,767; SAT-ELA sample size is 43,744

Table 2. Courses, Course GPA and Enrollment, Top 5% Courses

SCED Course	Course Name	Mean GPA	Enrolled
8051	Health Education	3.2	39850
8001	Physical Education	3.6	38488
3051	Biology	2.7	37717
1001	English/Language Arts I (9th grade)	2.8	36703
1002	English/Language Arts II (10th grade)	2.7	35229
24053	Spanish II	2.9	31838
3101	Chemistry	2.7	30062
24052	Spanish I	3.0	29764
1003	English/Language Arts III (11th grade)	2.5	27758
4101	U.S. History Comprehensive	2.8	24165
2056	Algebra II	2.6	23258
1004	English/Language Arts IV (12th grade)	2.6	22906
2072	Geometry	2.7	21377
4161	Civics	2.8	21299
4201	Economics	2.7	20140
4051	World History Overview	2.8	16567
24054	Spanish III	3.0	16292
2061	Integrated Math—multi-year equivalent	2.6	16229
2052	Algebra I	2.5	15401
3201	Integrated Science	2.6	15222
2110	Pre-Calculus	2.8	14225
4103	Modern U.S. History	2.8	12898
4001	World Geography	2.9	11903
3159	Physical Science	2.8	11782
22151	Career Exploration	3.5	11045
4254	Psychology	3.0	10993
3151	Physics	3.0	10824
8152	Drivers Education Class & Lab	3.0	8747
3053	Anatomy and Physiology	2.9	8343
2064	Integrated Mathematics III	2.4	7238
2062	Integrated Mathematics I	2.7	7072
3001	Earth Science	2.7	6981
1005	AP English Language and Composition	3.2	6704
4258	Sociology	3.1	6649
5154	Visual Arts Comprehensive	3.0	6615
2201	Probability and Statistics	2.7	6596
2063	Integrated Mathematics II	2.4	6381
8005	Fitness/Conditioning Activities	3.4	5866

Table 3. Average Course Difficulties by Content Area and AP/IB Designation

	Course Types	AP/IB	Math-AP/IB	Science-AP/IB	Arts-AP/IB	Humanities-AP/IB	Languages-AP/IB
	[1]	[2]	[3]	[4]	[5]	[6]	[7]
Science	0.078						
	(0.143)						
Arts	-1.005***						
	(0.137)						
Sports	-1.021***						
	(0.180)						
Humanities	-0.107						
	(0.126)						
Languages	0.059						
	(0.143)						
AP		1.395***	1.748***	1.588***	0.973***	1.297***	0.829***
		(0.137)	(0.509)	(0.214)	(0.281)	(0.194)	(0.271)
IB		1.117***	1.410***	1.299***	1.372***	0.246	1.101***
		(0.168)	(0.517)	(0.282)	(0.358)	(0.271)	(0.273)
Constant	-0.119	-0.576***	-0.276**	-0.375***	-1.259***	-0.389***	-0.247***
	(0.102)	(0.042)	(0.119)	(0.086)	(0.074)	(0.067)	(0.085)
Observations	411	411	62	63	80	112	65

Note: Reference category is Math in [1], Non-AP/IB in [2], and Non-AP/IB Subject Area in [3]--[7]. Estimates based on meta-analytic regressions, inverse variance weighted by the standard error of the course difficulty. * 0.1 ** 0.05 *** 0.01

Table 4. SAT and $\widehat{\theta}$ Alignment and Departure: Transcript Case Study

		High SAT - High Theta		**High SAT - Low Theta**		**Low SAT - High Theta**		**Low SAT - Low Theta**	
		Math=583;ELA=595; Theta=2.0;N=12,477		Math=531;ELA=552; Theta=-1.2;N=2,067		Math=405;ELA=422; Theta=0.8;N=1,867		Math=381;ELA=391; Theta=-1.6;N=12,302	
	Difficulty	GPA	% Taken	GPA	% Taken	GPA	% Taken	GPA	% Taken
Physical Education	-1.63	3.94	90%	3.42	88%	3.82	86%	3.05	86%
Health Education	-1.03	3.84	93%	2.82	91%	3.70	87%	2.50	89%
Spanish I	-0.79	3.77	59%	2.44	62%	3.59	72%	2.17	75%
U.S. History	-0.59	3.71	46%	2.38	58%	3.36	51%	1.94	63%
Spanish II	-0.58	3.72	70%	2.28	69%	3.52	74%	2.12	74%
ELA III (11th grade)	-0.54	3.60	36%	2.03	70%	3.35	70%	1.82	83%
ELA I (9th grade)	-0.53	3.62	79%	2.25	86%	3.25	82%	1.97	87%
ELA II (10th grade)	-0.50	3.60	70%	2.16	84%	3.35	79%	1.98	88%
Biology	-0.44	3.65	83%	2.26	85%	3.20	87%	1.86	88%
Algebra II	-0.26	3.51	52%	2.04	51%	3.25	56%	1.76	52%
Chemistry	-0.16	3.52	78%	1.95	72%	3.16	70%	1.72	55%
Pre-Calculus	0.50	3.37	55%	1.79	36%	2.92	31%	1.81	8%
AP Psychology	0.58	3.47	25%	1.72	11%	2.99	8%	1.66	2%
AP English Language and Comp.	0.82	3.54	36%	1.84	9%	3.20	8%	1.94	2%
AP English Literature and Comp.	0.86	3.54	21%	1.72	5%	3.08	5%	1.77	1%
AP U.S. History	0.87	3.48	32%	1.62	9%	2.94	6%	1.40	1%
AP U.S. Govt and Politics	0.90	3.43	11%	1.63	4%	3.09	2%	1.70	1%
Calculus	0.97	3.35	24%	1.88	8%	2.94	3%	2.08	0%
AP Biology	1.03	3.48	21%	1.60	5%	3.05	3%	1.59	1%
AP Statistics	1.18	3.42	24%	1.56	5%	2.97	2%	1.65	0%
AP Chemistry	1.39	3.51	13%	0.90	1%	3.18	1%	1.29	0%
AP Calculus AB	1.58	3.33	24%	1.35	4%	2.97	2%	1.88	0%

Note: Each group is based on the top two quintiles of standardized average Math and ELA SAT and Theta. We define quadrants as (i) highest two quintiles SAT and Theta; (ii) highest two quintiles SAT and lowest two quintiles of Theta; (iii) lowest two quintiles SAT and highest two quintiles of Theta; (iv) lowest two quintiles of SAT and Theta. Quadrants are not equally representative of high and low scoring SAT/Theta students because these two variables are correlated. High/high and low/low are more comprised of higher (lower) performing students, whereas high/low and low/high are more likely to represent the average student. In our sample, about 100 points in the SAT scale corresponds to 1 standard deviation meaning that high/low SAT groups are nearly 2 full standard deviation apart. 1.9 points of the Theta scale corresponds to 1 standard deviation meaning that high/low Theta groups are about 1.6 standard deviations apart. Courses presented are limited to those with at least 50% of students having taken (among students in these four quadrants), or an AP course with at least 5% participation, or a Pre-Calculus or Calculus course.

Table 5. Relationship between GPA, $\hat{\theta}$, and Mean Math and ELA SAT Scores on College-Going Behavior

	Average Graduate Earnings		College Tier		2- vs 4-Year Degree	
Panel A: GPA and Theta Separately						
GPA	6291.18***		2.37***		2.85***	
	(94.94)		(0.03)		(0.09)	
SAT	6128.41***	5131.77***	1.91***	1.75***	1.66***	1.48***
	(94.91)	(102.34)	(0.02)	(0.02)	(0.05)	(0.04)
Theta		7148.40***		2.50***		3.21***
		(102.51)		(0.04)		(0.11)
N	41286		43767		8395	
Theta = GPA	0.000		0.000		0.000	
SAT$_1$ = SAT$_2$	0.000		0.000		0.000	
Panel B: Theta Conditional on GPA {2.0, 2.5, 3.0, 3.5}						
Theta	28979.89***		31.46***		123.80***	
	(3904.82)		(17.44)		(159.09)	
SAT	3566.13***		1.53***		1.10	
	(503.87)		(0.10)		(0.16)	
N	2015		2173		417	
Theta = SAT	0.000		0.000		0.001	
Panel C: Theta Conditional on Rounded GPA {1.0(0.1)4.}						
Theta	25235.01***		21.88***		20.52***	
	(813.30)		(2.30)		(5.50)	
SAT	4209.41***		1.60***		1.36***	
	(111.68)		(0.02)		(0.04)	
N	41285		43767		8395	
Theta = SAT	0.000		0.000		0.000	

Note: *Outcomes* - Earnings data are from the Chetty College Report Card and represent median child earnings in 2014 for the 1980-1982 birth cohorts, including non-college attendees. College Tier is an ordinal variable from 1-5 representing highly selective, selective, non-selective, two-year, and no college. 2- vs 4-Year College Degree is an ordinal variable indicating whether a student completed a 4-year degree, 2-year degree, No degree but some college, or no college. Degree completion is restricted to the 2017 graduating cohort. For both ordinal outcomes, coefficients are reported as odds ratios. *Models* - In Panel A, GPA and Theta scores are standardized and estimated in separate models. We then stack the parameter estimates with their variance/covariance matrices to test coefficient equality across models. The p-values report two tests: (1) whether Theta and GPA coefficients are equal, and (2) whether SAT coefficients differ between Models 1 and 2. In Panel B, we include GPA fixed effects for GPAs {2.0(0.5)3.5} following results in Figure 2. The p-value reports whether SAT and Theta coefficients are equal. In Panel C, we include GPA fixed effects for rounded GPA scores {1.0(0.1}4.0}. Robust standard errors are shown in parentheses. * 0.1 ** 0.05 *** 0.01

Table 6. Comparing GPA and $\widehat{\theta}$ across three simulation studies.

	Study 1	Study 2	Study 3
$\hat{\theta}$ based on observed			
Mean bias	0.006	0.027	-0.006
Mean RMSE	0.463	0.376	1.174
Mean NRMSE	0.130	0.162	0.513
Mean CCC	0.967	0.979	0.724
Observed GPA			
Mean bias	0.104	0.181	0.168
Mean RMSE	0.46	0.439	0.523
Mean NRMSE	0.832	0.493	0.635
Mean CCC	0.456	0.815	0.724
Correlations			
Mean $\hat{\theta}$ - GPA correlation	0.738	0.967	0.963
SD $\hat{\theta}$ - GPA correlation	0.046	0.004	0.007
NRMSE compared to θ			
Full data $\hat{\theta}$	0.068	0.078	0.08
Full data GPA	0.094	0.239	0.243

Note. Bias = mean signed difference between estimate and criterion. RMSE = root-mean-squared difference between estimate and criterion. NRMSE = RMSE computed with variables standardized. CCC = concordance correlation coefficient (Lin, 1989). Criterion for $\hat{\theta}$ based on observed data is $\hat{\theta}$ based on full data. Criterion for observed GPA is GPA based on full data. Reported correlations are for values of $\hat{\theta}$ and GPA based on observed data. NRMSE compared to θ = NRMSE relative to data-generating true θ value. These comparisons are based on the full-data $\hat{\theta}$ and GPA.

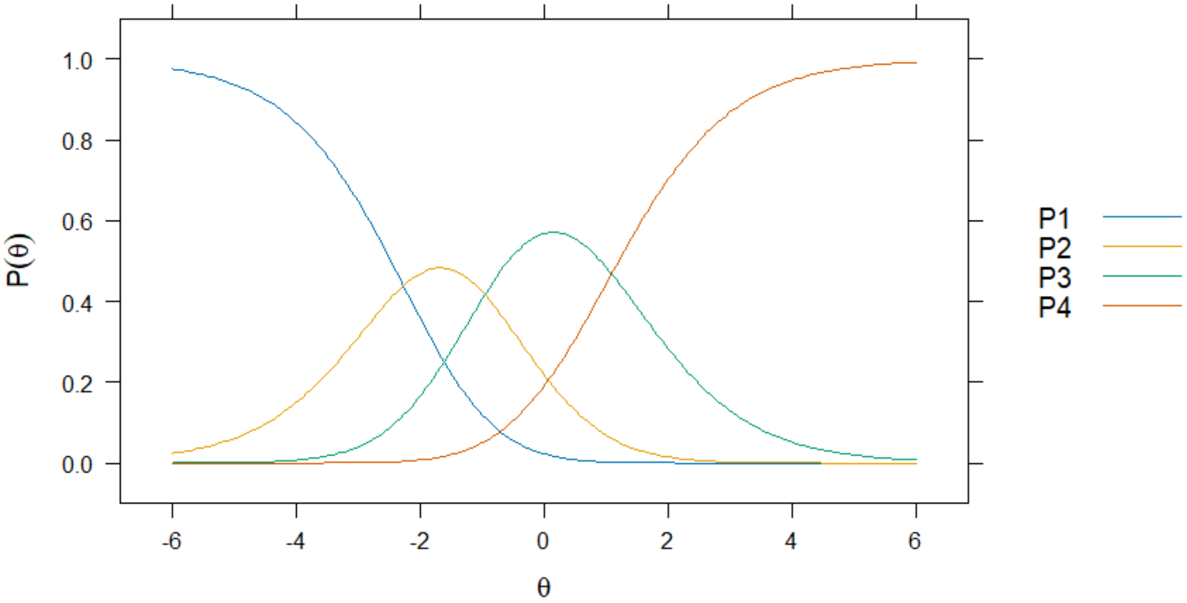
Table 7. Standardized Differences in $\widehat{\theta}$, SAT Scores, and Theta Conditional on SAT, by Subgroup

	Theta Standardized	Mean SAT Standardized	Theta | SAT
Panel A: Gender			
Male (vs Female)	-0.413***	-0.062***	-0.369***
	(0.009)	(0.010)	(0.006)
Constant	0.204***	0.031***	0.182***
	(0.007)	(0.006)	(0.005)
Panel B: Race/Ethnicity			
Asian (vs White)	0.678***	0.687***	0.202***
	(0.027)	(0.030)	(0.020)
Black (vs White)	-0.616***	-0.763***	-0.088***
	(0.010)	(0.010)	(0.008)
Hispanic (vs White)	-0.440***	-0.651***	0.010
	(0.013)	(0.012)	(0.010)
Other (vs White)	-0.244***	-0.253***	-0.069***
	(0.030)	(0.030)	(0.022)
Constant	0.230***	0.306***	0.018***
	(0.007)	(0.007)	(0.005)
Panel C: Income			
Low Income (vs Not)	-0.618***	-0.633***	-0.183***
	(0.009)	(0.009)	(0.007)
Constant	0.208***	0.214***	0.062***
	(0.006)	(0.006)	(0.004)
Panel D: SWD			
SWD (vs Not)	-0.748***	-1.039***	-0.010
	(0.010)	(0.011)	(0.010)
Constant	0.095***	0.132***	0.001
	(0.005)	(0.005)	(0.004)
Panel E: ELL			
ELL (vs Not)	-0.158***	-0.389***	0.121***
	(0.013)	(0.013)	(0.010)
Constant	0.021***	0.051***	-0.016***
	(0.005)	(0.005)	(0.004)
N	43767	43767	43767

Note: OLS regression coefficients reported; all models include graduating year fixed effects. Robust standard errors are shown in parentheses. Panels A--E represent separate sets of regression models to avoid controlling for group-specific factors separate from the subgroups under investigation (e.g., to avoid controlling for socioeconomic status in regressions testing for differences by race/ethnicity). * 0.1 ** 0.05 *** 0.01

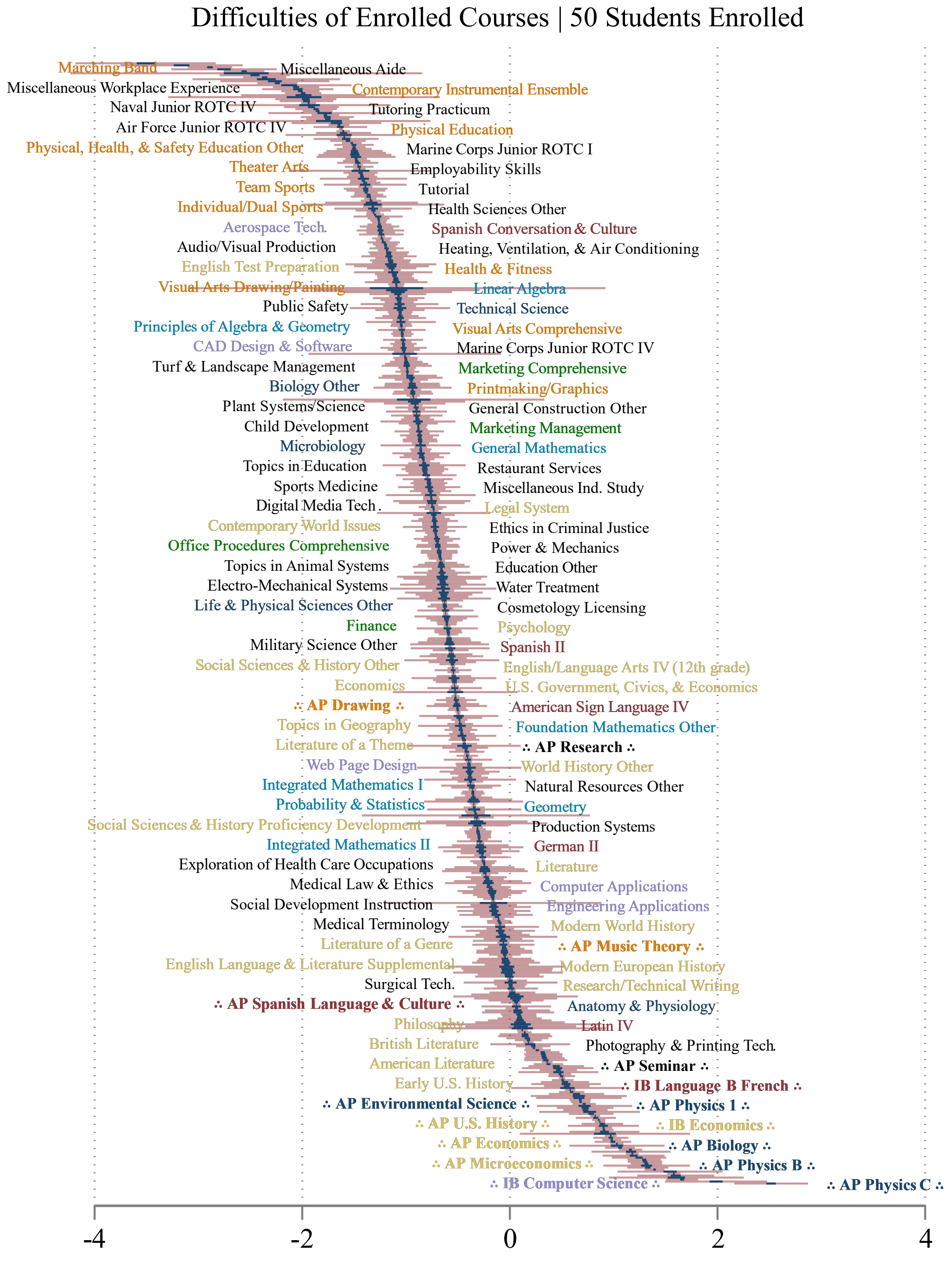
Figures

Can Figure 1. Example Partial Credit Model Item Characteristic Curve for an Item with Four Response Categories



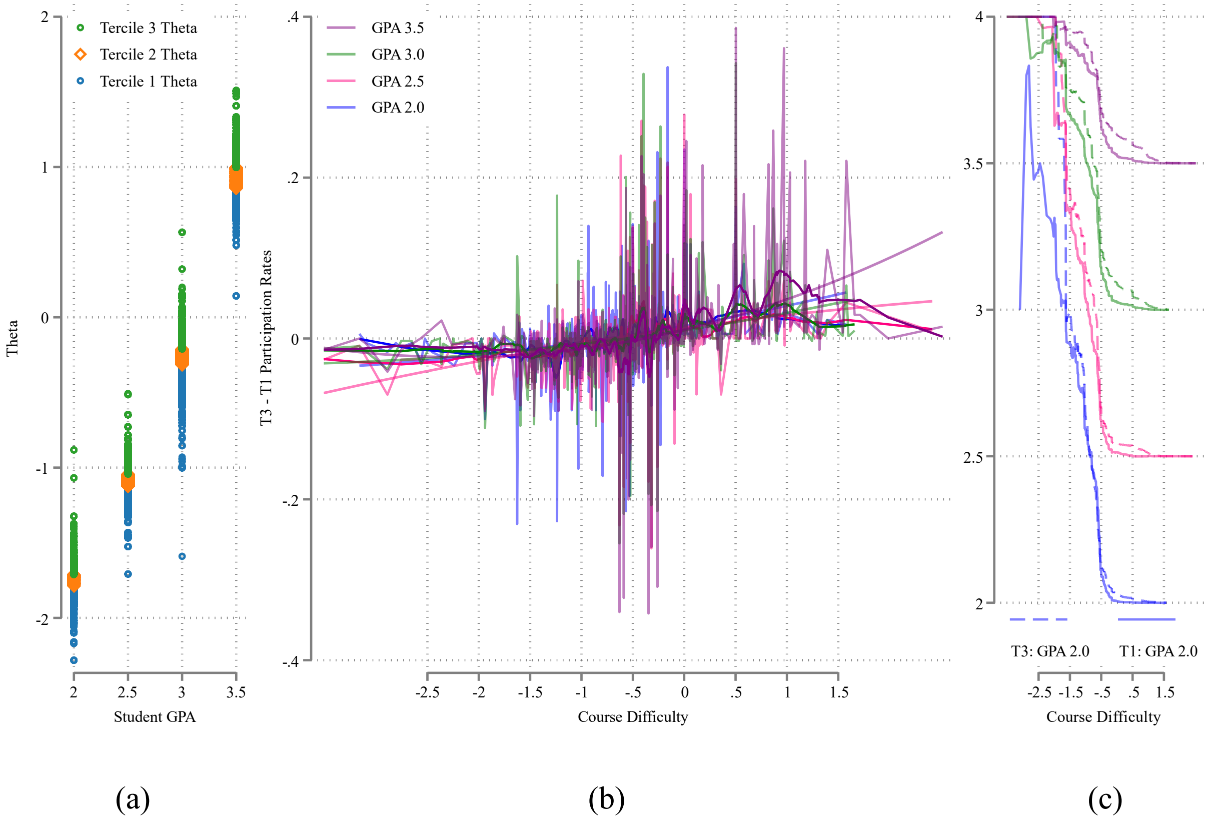
Note: Figure depicts an example Item Characteristic Curve for the Partial Credit Model (Masters, 1982). The curves depict on the y-axis the modeled probability of responding in each of four ordered response categories, P1-P4, according to the estimated parameters of the model for a single item. The model's threshold parameters are the points at which the curves for each adjacent pair of categories cross. The x-axis is the person ability continuum θ, such that probabilities of responding in each category for a given value of θ can be computed.

Figure 2. Meta-Analytic Average Difficulty by Course



Note: Courses sorted by difficulty. Every 10th course name is labeled. AP and IB courses are bold and have three-dot symbol adjacent to course name. Course name colors indicate content: math is colored light blue, science is colored dark blue, computer courses colored lavender, humanities colored tan, arts or sports colored orange, languages colored maroon, and business colored green.

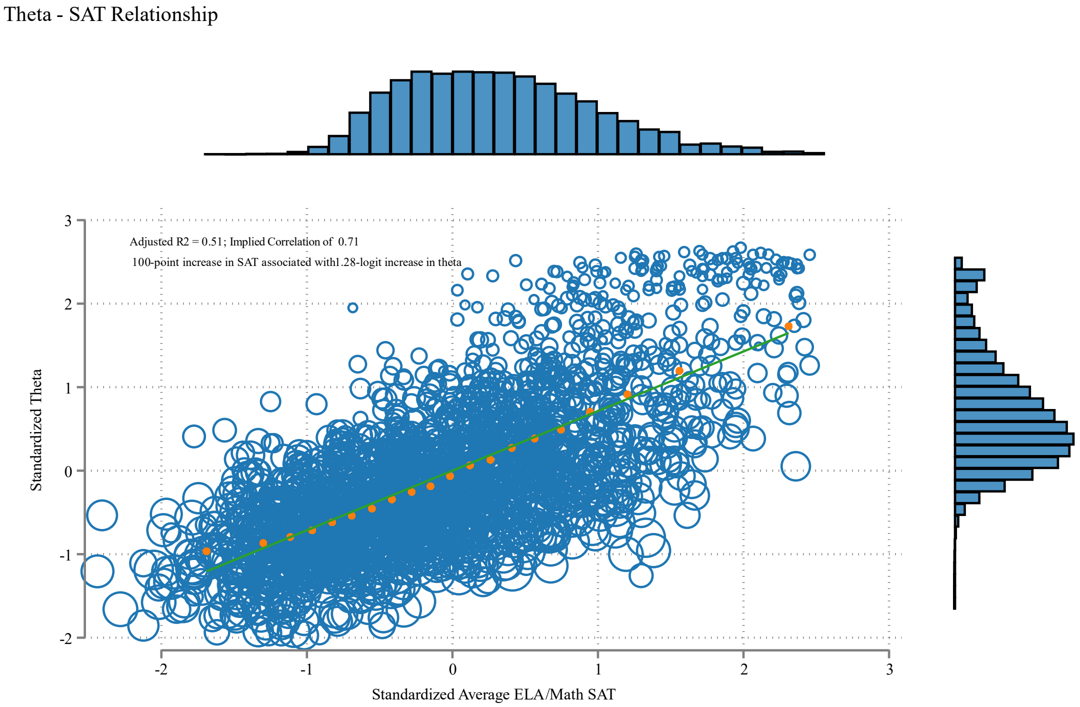
Figure 3. GPA-$\hat{\theta}$Relationship



(a) (b) (c)

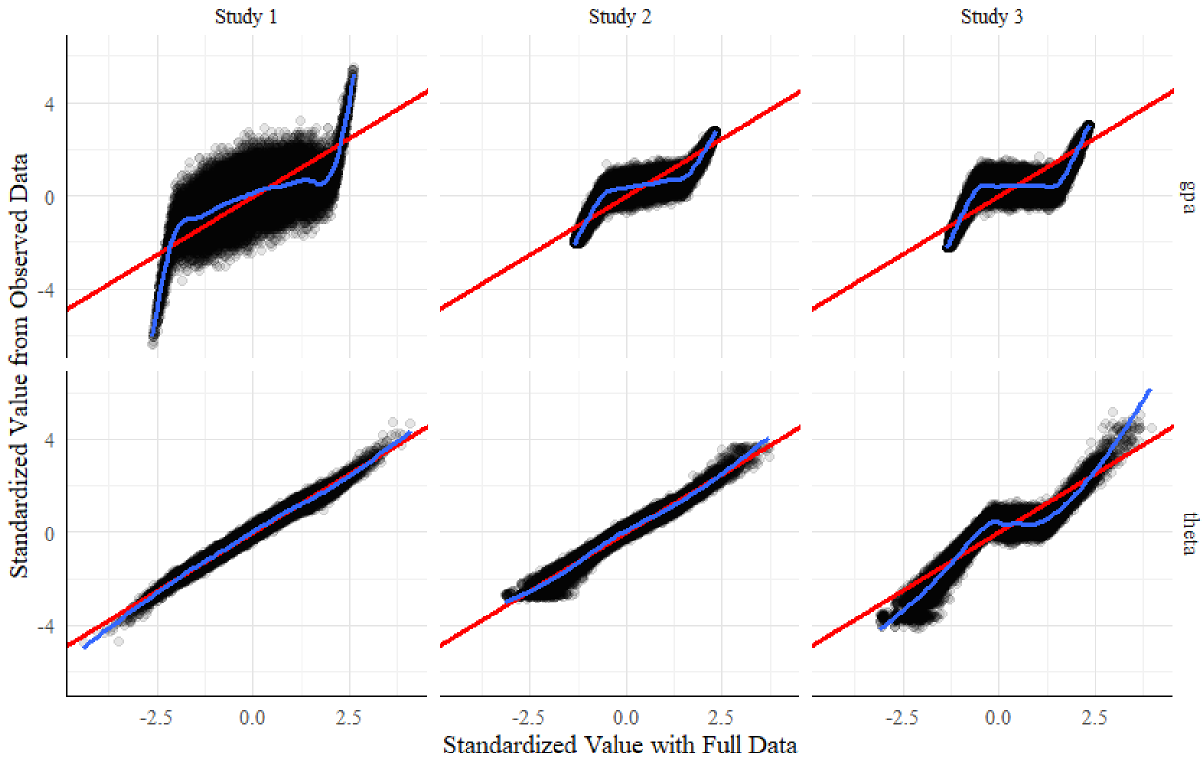
Note: Panel (a) shows the variation in Theta for 4 student-level GPAs – 2.0 (N=515), 2.5 (N=332), 3.0 (N=912), and 3.5 (N=414). For each GPA level, we split Theta into three terciles, showing substantial variation in true ability (Theta) even among students who achieve the same GPA. Panel (b) shows the difference in course-level participation rates between Tercile 3 and Tercile 1 Theta students for each course and GPA score as a function of course difficulty. The thin lines are raw data, medium lines are lowess lines estimated with a 0.05 bandwidth, and the thick lines are quadratic fits. Positive slopes indicate that Tercile 3 students are more likely to take high difficulty courses than Tercile 1 students among students with the same GPA. This consistent pattern across all GPA levels shows that higher-ability students systematically (albeit modestly) select into more challenging courses regardless of their GPA level. Given equivalent GPA levels between tercile groups (high-low theta), this suggests that selection into more difficult courses reduces GPA for Tercile 3 students. This result is confirmed in Panel (c), which shows how course selection patterns contribute to final GPAs. Starting from any point on the difficulty scale, taking additional difficult courses tends to lower cumulative GPA. While Tercile 3 (higher ability) students maintain higher GPAs through easier courses, their systematic selection into more difficult courses ultimately brings their GPAs in line with Tercile 1 students, explaining the variation in Theta within GPA bands shown in Panel (a).

Figure 4. $\widehat{\theta}$-SAT Relationship



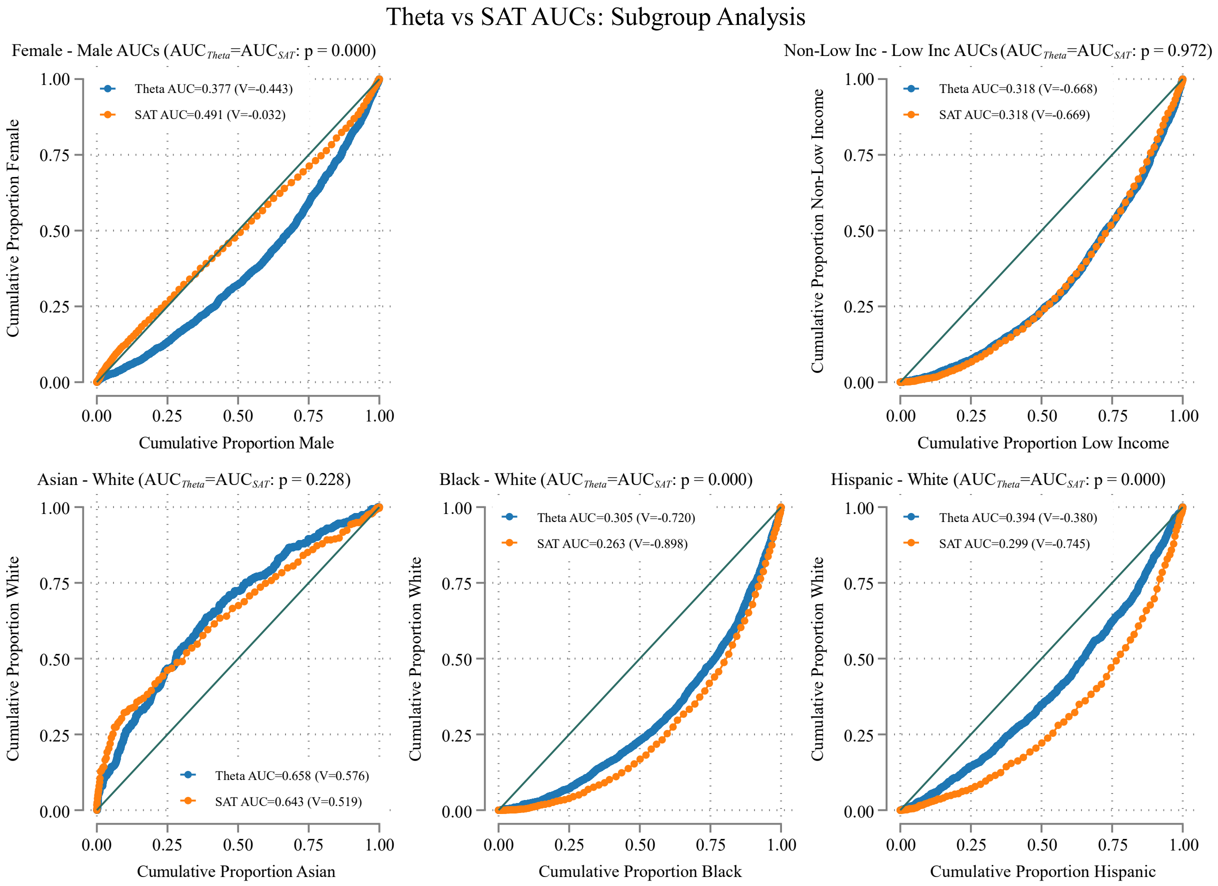
Note: Scatter Plot of 2000 randomly selected students' standardized Theta and Mean Math & ELA SAT scores; fitted line based on full sample. Histograms of the two measurements are shown alongside the Y- and X-axes.

Figure 5. Comparing Observed GPA and $\widehat{\theta}$ from 28 Selected Courses to Equivalents from 200 Courses without Missingness



Note: Red line = identity line. Blue curve = generalized additive model-based smoothed curve. X-axis is based upon computing GPA or estimating $\widehat{\theta}$ using simulated responses to all 200 courses in simulation. Y-axis is based upon computing GPA/estimating $\widehat{\theta}$ from 28 courses selected from 200 total via ability-based selection mechanism. Study 1: no range restriction on course difficulty, includes “core” of 8 courses of which each examinee takes five. Study 2: range restriction on course difficulty, includes same core as study 1. Study 3: range restriction on course difficulty, with no core.

Figure 6. Ordinal Gaps for Various Subgroups in the $\widehat{\theta}$and SAT Metrics



Note: AUC estimated via Rocfit following (Ho, 2009 and Reardon & Ho, 2012) and converted to a V-statistic. AUC is interpreted as the probability that a randomly selected student in group *g1* (e.g., male) in Theta or SAT metric scores higher than randomly selected student in group *g2* (e.g., female). V-statistic can be interpreted as standardized mean difference between the groups of students under assumption of respective normality. The difference in the two AUCs is conducted using the Delong, et al. (1988) test and the p-value of that test is reported in each figure panel title. .

Technical Appendices

Appendix A: IRT Model Fit Results

Model Results

The model converged after 516 iterations. Here, we present several pieces of information summarizing the results of model fitting in terms of the measurement precision of θ as well as the appropriateness of the model for the data.

Person and item reliability

One of the most important considerations when using student-level scores in quantitative research is the reliability of the scores – the proportion of variance in the scores attributable to "true" underlying variance in the construct of measurement, rather than random measurement error. The notion of reliability has its roots in classical test theory (CTT) and is often described in reference to the canonical CTT formula $X = T + E$, referring to observed scores (*X*) being the sum of the true score *T* and an uncorrelated random error *E*. Reliability in CTT, given the conceptual definition above, would be equal to $1 - \sigma_E^2 / \sigma_x^2$. While this cannot be computed directly given that the terms in $X = T + E$ are unknowable, many approaches to estimating reliability exist (Cronbach, 1951; Guttman, 1945; etc.).

The CTT conceptualization implies a constant error term that is the same for all individuals, but in IRT measurement precision, quantified as the standard error of $\hat{\theta}$, changes for different values of $\hat{\theta}$ and, in our case, as a function of how many and which courses a student took (more courses, and the difficulty of those courses being closer to the student's ability, will produce a smaller standard error). Still, it is critically important to know that when we order students by $\hat{\theta}$, we are ordering them on a variable that is not overly error-laden; an excess of measurement error can lead to lower power to detect group-level differences, as well as create false positives in multiple regression (Shear & Zumbo, 2013). To that end, we report what Chalmers (2012) refers to as the empirical reliability of $\hat{\theta}$, r_{xx}, calculated as:

$$r_{xx} = 1 - \frac{\overline{SEM}^2}{\sigma_{\hat{\theta}}^2 + \overline{SEM^2}} = \frac{\sigma_{\hat{\theta}}^2}{\sigma_{\hat{\theta}}^2 + \overline{SEM^2}} \tag{A2}$$

where $\sigma_{\hat{\theta}}^2$ is the variance of $\hat{\theta}$ and $\overline{SEM}$ is the mean standard error of $\hat{\theta}$. This is very similar to the CTT conceptualization of reliability; to get the equivalent of E, we marginalize over the many unique values of the standard error. For our measurement model and population, r_{xx} is equal to 0.948, indicating that ~95% of the variance in $\hat{\theta}$ is not attributable to random measurement error. Given that 0.7 is often cited as a minimum reliability to use scores from an instrument in research (Bandalos, 2018), our measure more than exceeds common standards for reliability.

Part of our analysis of the results of this model fitting exercise below involves ordering courses by their difficulty (average threshold) below. Just as score reliability is important when using scores, so too is the reliability of the difficulties quite important when one wants to order items by their difficulty. To that end, we also report the reliability of our threshold parameters, representing the proportion of variance in the item parameters not attributable to error. Here, empirical reliability is 0.879, again indicating that error is not adding an undue amount of noise into our analyses of course difficulty below. However, as detailed below, we use an inverse variance-weighted mean threshold to represent course difficulty for the purpose of ordering courses, based upon the observation that thresholds are estimated with widely varying error variances as a function of course-specific grade distributions and sample sizes.

Person and item-side parameter distributions

One of the most common ways of making sense of the relationship between item difficulty and person ability when using Rasch-type IRT models is a plot often referred to as a Wright Map (these are described in detail in e.g. Bond & Fox, 2015; Wilson, 2023). A Wright Map plots item difficulties as individual points, and the $\hat{\theta}$ distribution as a histogram, on opposite sides of the same axis representing the logit scale (recall that ability and difficulty estimates are on a common scale). A Wright Map is not feasible for the main model used in this study because of the number of courses in the model (742 producing 2549 total threshold parameters). We therefore present a slightly modified Wright Map below in Figure A1. Here, we present parallel plots of the $\hat{\theta}$ distribution and the distribution of threshold parameters, color-coded by threshold. This figure indicates two key findings underscoring the plausibility of the measurement model. First, the spread of the two distributions is similar, indicating that there are courses in which even students who struggle the most academically are quite likely to earn at least a C, and courses in which all but the most academically successful struggle to earn an A. Second, the color-coding of the threshold parameters indicates that the thresholds are generally ordered as expected in the aggregate (that is, first thresholds are generally easier to pass than second thresholds, which are in turn easier than the third); this is not guaranteed by the model itself (Andrich, 2005), but is desirable for the task of ordering and interpreting the difficulties of items.

That said, 45.3% of the courses used in our analytic sample did have at least one disordered threshold, meaning that $\tau_{i1} > \tau_{i2}$,$\tau_{i2} > \tau_{i3}$, or $\tau_{i3} > \tau_{i4}$. It is important to emphasize that a disordered threshold does not typically indicate that the higher grade in that course is, in fact, *easier* to achieve than the lower grade. Rather, disordered thresholds are often a result of responses in one of the categories for which there is a disordered threshold being very infrequent. To illustrate how a disordered threshold can result from low response frequencies in a middle category, we present Figure A2. This shows the ICC for one of the courses in our model that had disordered thresholds. As is the case for most disordered thresholds, the disordering is a function of low frequency of the second-lowest response category (a course grade of D). In fact, when one limits the inspection of threshold ordering to just τ_{i2} through τ_{i4}, the percentage of courses with a disordered threshold is just 11.0%. Given our intent to represent the data as an external reader would see it via the PCM, we opted not to recode any course grades, and to proceed with some disordering of the first threshold. Ultimately, we note the existence of these disordered thresholds, but do not consider them problematic for measurement, in line with Adams et al. (2012).

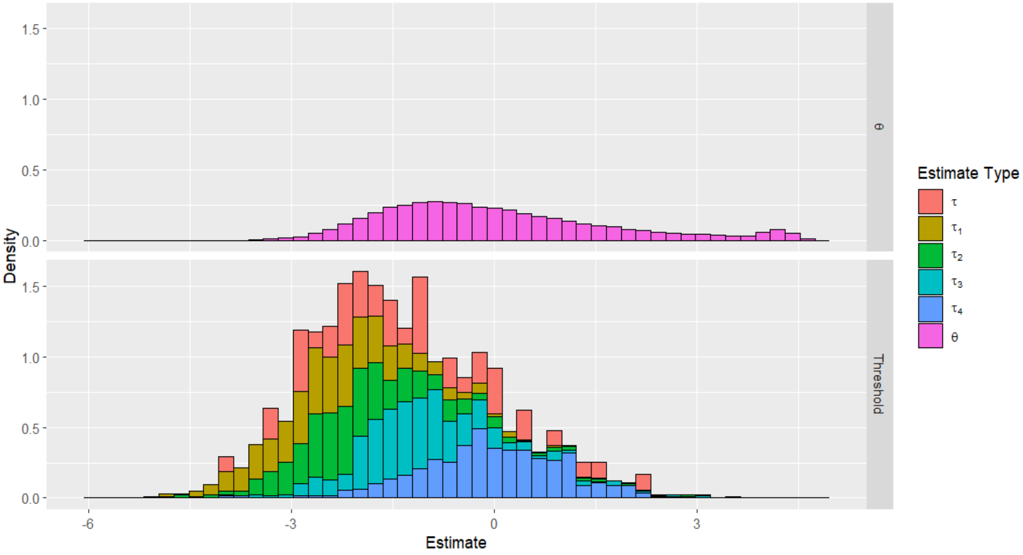


Figure A1. Score and Item Parameter Distributions.

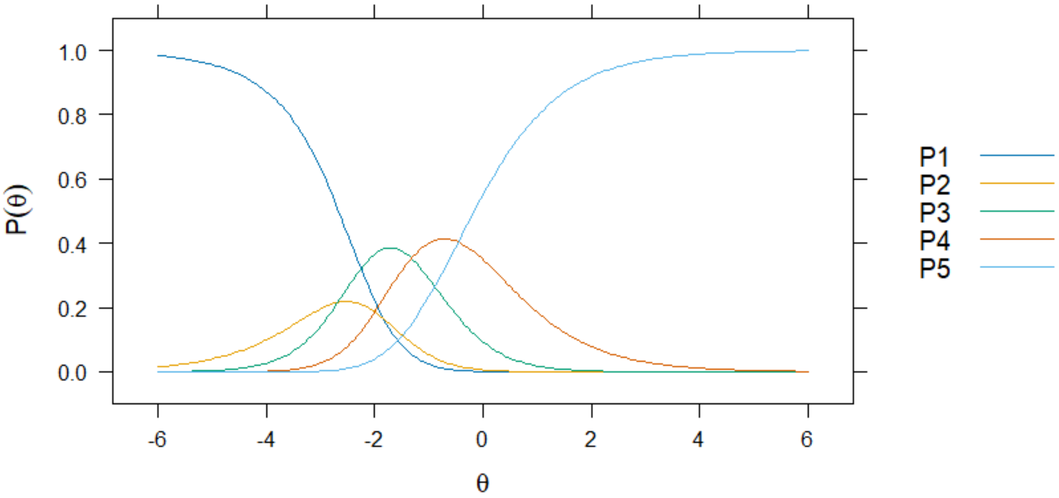


Figure A2. Example Partial Credit Model Item Characteristic Curve for an Item with Disordered Thresholds Due to Low Response Frequency

Model-data fit

When using Rasch-type models, it is quite important to ensure that the model is appropriate for the data, a task typically approached at the item level via item fit statistics (Bond & Fox, 2015; Wu & Adams, 2013). We investigated model-data fit from two angles. First, we computed and reviewed the mean-square fit statistic known as information-weighted fit, or "infit" in the Rasch tradition. Briefly, infit is a fit statistic based upon mean squared standardized residuals of item responses, with an expected value of 1 where values above 1 indicate under-fit to the measurement model (i.e. poor fit) and values below 1 indicate over-fit to the model (i.e. data that are too predictable and do not contribute much information to the model). Values above 1.5 are generally considered concerning (Linacre, 2002), while values below 1, no matter how low, are generally not considered an issue in secondary data analysis where it does not make sense or is not possible to add or remove items; such items are inefficient from a measurement perspective but not considered problematic (Wright & Linacre, 1994). Of the 742 courses, only 19 had infit values above 1.5, indicating that the model is generally appropriate for the data. An unweighted version of infit that is more sensitive to outliers, "outfit" (for outright or outlier-sensitive fit), was also computed, with 47 courses producing values above 1.5. In short, the data appear to fit the PCM well.

However, traditional Rasch fit statistics have been criticized as being sample size-dependent (Müller, 2020; Wu & Adams, 2013), and their practical significance is not always clear, especially in our dataset where different courses were taken by widely differing numbers of students and the number of items is far larger than one would find on a typical educational test. To validate our claim from our outfit analysis that item fit is generally acceptable, we also use a plausible values imputation-based fit analysis (Chalmers & Ng, 2017 introduce and outline this approach; for brevity, we refer readers to this paper) to assess the practical significance of whatever item misfit does exist in our data. The analysis is based on the Yen's Q_1 fit statistic (Yen, 1981), with plausible values imputations used to account for measurement error in $\hat{\theta}$. As implemented in *mirt*, this procedure produces a key measure of the practical significance of item misfit, an item-specific root mean square error of approximation (RMSEA) value. Following prior work using RMSEA to assess item-level practical significance of misfit (Oliveri & Von Davier, 2011), we consider a value of 0.1 or higher to be cause for concern. By this measure, just 23 courses misfit the model at a practically significant level (and only 26 misfit at the more stringent cutoff of 0.05), indicating that misfit is likely negligible for the purposes of analyzing item parameters and $\hat{\theta}$s.

Sensitivity to population and modeling choices

To devise the approach outlined above, several times we had to choose between two (or more) approaches to a question with no clear answer. Specifically, we note the choices to (1) take the mean score from multiple course attempts, (2) exclude students with no SAT score from the model fitting process, and (3) use the somewhat more restrictive PCM instead of the more flexible Generalized Partial Credit Model (Muraki, 1992) or Graded Response Model (Samejima, 1969). As a sensitivity check, we fit these alternate IRT models and changed our rule for multiple courses to keep only the highest grade. We also ran our main model including students without SAT scores. We did so to assess the extent to which these choices impacted the $\hat{\theta}$ scores that we analyze in the main study. We found Pearson correlations between all sets of $\hat{\theta}$ that were above 0.98, indicating that even if we had made different choices in our measurement procedure, we likely would have arrived at very similar results. The more complex generalized PCM and Graded Response Model also produced very unstable (i.e. large standard errors) discrimination parameters for many of the courses (mainly those with lower enrollment), calling into question the value of the more flexible models.

References

Adams, R. J., Wu, M. L., & Wilson, M. (2012). The Rasch Rating Model and the Disordered Threshold Controversy. *Educational and Psychological Measurement*, *72*(4), 547–573. https://doi.org/10.1177/0013164411432166

Andrich, D. (2005). The Rasch Model explained. In S. Alagumalai & J. P. Keeves (Eds.), *Applied Rasch measurement: A book of exemplars* (pp. 27–59). Springer.

Bandalos, D. L. (2018). Measurement theory and applications for the social sciences. Guilford Press.

Bond, T. G., & Fox, C. M. (2015). Applying the Rasch Model: Fundamental Measurement in the Human Sciences (3rd ed.). Routledge.

Chalmers, R. P., & Ng, V. (2017). Plausible-value imputation statistics for detecting item misfit. *Applied Psychological Measurement*, *41*(5), 372–387. https://doi.org/10.1177/0146621617692079

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *15*(3), 298–334.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*(4), 255–282. https://doi.org/10.1007/BF02288892

Linacre, J. M. (2002). What do Infit and Outfit, Mean-square and Standardized mean? *Rasch Measurement Transactions*, *16*(2), 878.

Müller, M. (2020). Item fit statistics for Rasch analysis: Can we trust them? *Journal of Statistical Distributions and Applications*, *7*(1), 5. https://doi.org/10.1186/s40488-020-00108-7

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *ETS Research Report Series*. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x

Oliveri, M. E., & Von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling*, *53*(3), 315–333.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, *34*(2, pt. 2), 100.

Shear, B. R., & Zumbo, B. D. (2013). False positives in multiple regression: Unanticipated consequences of measurement error in the predictor variables. *Educational and Psychological Measurement*, *73*(5), 733–756. https://doi.org/10.1177/0013164413487738

Tyner, A., & Gershenson, S. (2020). Conceptualizing grade inflation. *Economics of Education Review*, *78*, 102037.

Wilson, M. (2023). Constructing measures: An item response modeling approach (2nd ed.). Routledge.

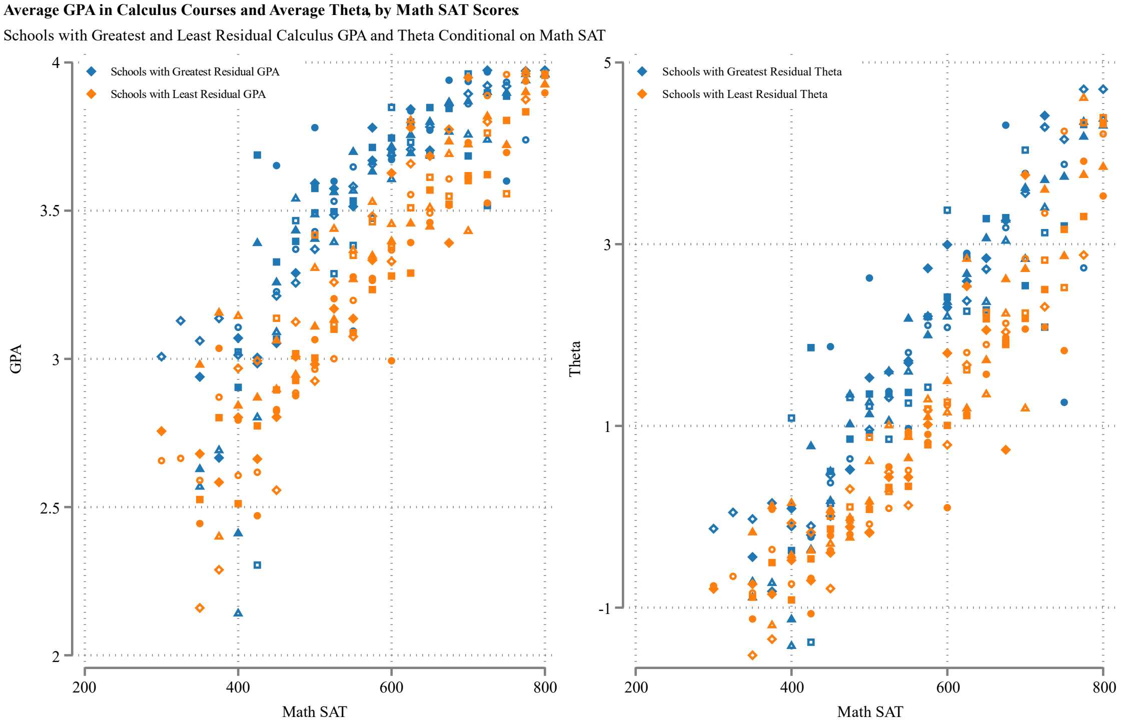
Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370.

Wu, M., & Adams, R. J. (2013). Properties of Rasch residual fit statistics. *Journal of Applied Measurement*, *14*(4), 339–355.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, *5*(2), 245–262. https://doi.org/0146-6216/81/020245-18$1.90

Appendix B: Grade Inflation

Figure B1. Average School-Level GPA in Calculus Courses and Average Theta, by Math SAT Scores



Note: This figure presents school-level variation in Calculus performance and estimated student ability. Each panel plots outcomes against Math SAT scores for two groups of schools - those with the highest (blue) and lowest (orange) residual variation in Calculus GPA after controlling for Math SAT, with markers representing individual schools. The left panel shows mean Calculus GPA by school, while the right panel displays mean estimated $\hat{\theta}$ (latent transcript strength) for the same schools. The systematic gap between high- and low-residual schools at similar Math SAT levels raises the possibility of institutional variation in grading standards that is not explained by differences in student ability as measured by the SAT. This pattern persists, though moderates, when examining $\hat{\theta}$.

Estimation Procedure for Course-Level Grade Inflation

We estimate course-level grade inflation using a leave-one-out procedure that prevents each school's own grading practices from influencing its estimated inflation measure. For each school *j* and course *c* combination, we estimate the following regression using data from all schools except school *j*:

$$y_{isc} = \beta_0 + \beta_1 SAT_i^{ELA} + \beta_2 SAT_i^{MATH} + \varepsilon_{isc}$$

where:

- y_{isc} is the course grade for student *i* in school *s* and course *c*
- SAT_i^{ELA} and SAT_i^{MATH} are the student's SAT verbal and mathematics scores
- ε_{isc} is the error term

For each student *i* in school *j* and course *c*, we then compute:

1) The predicted grade: $\hat{y}_{ijc} = \hat{\beta}_0 + \hat{\beta}_1 SAT_i^{ELA} + \hat{\beta}_2 SAT_i^{MATH}$
2) The inflation measure: $\gamma_{ijc} = y_{ijc} - \hat{y}_{ijc}$
3) This approach yields a school-course specific measure of grade inflation that:
   a) Controls for student preparation as measured by SAT scores
   b) Avoids mechanical correlation with a school's own grading practices through the leave-one-out design
   c) Allows for course-specific relationships between student preparation and expected performance

The resulting γ_{ijc} represents the degree to which grades in course *c* at school *j* deviate from what would be predicted based on student SAT scores, using the relationship between SAT scores and grades established at other schools.

We then examine how our estimated measure of grade inflation or a student's mean SAT score more strongly relates to latent transcript strength ($\hat{\theta}$) through a series of regressions, adding additional fixed effects (results in Table A1). The base specification is:

$$\hat{\theta}_i = \alpha + \beta_1 I_i + \beta_2 S_i + \gamma_c + \varepsilon_i$$

where $\hat{\theta}_i$ is the standardized Theta score for student i, I_i is their standardized grade inflation exposure, S_i is the mean SAT score, and γ_c is graduation cohort year fixed effect. We estimate models with inflation alone and then add in the SAT score (Table A1, columns [1] and [2]).

Then, we include school fixed effects (Table A1, columns [3] and [4]) to test whether the relative influence of grade inflation is changed when controlling for between school differences. Then, we include school-by-course fixed effects (Table A1, columns [5] and [6]) to test whether the relative influence of grade inflation is changed when controlling for between school-by-course differences.

The results reveal several key findings:

1. Grade inflation consistently predicts Theta with coefficients ranging from 0.348 to 0.392 standard deviations
2. SAT scores show stronger predictive power, with coefficients between 0.694 and 0.735 standard deviations
3. Semi-partial correlations indicate that SAT scores explain more unique variance in Theta than grade inflation (0.321-0.432 vs 0.137-0.218)
4. The relationship between grade inflation and Theta remains robust across all specifications, including our most stringent model with school-by-course fixed effects, suggesting that cross school and cross school-course grading practices are not responsible for the influence of grade inflation on theta.

This analysis provides evidence that grade inflation is likely to have some influence on latent transcript strength, an independent measure of student ability via the SAT more closely tracks the quality of a student's coursetaking history.

Table B1. College Outcomes for Top/Bottom Quintiles of Theta / SAT Distribution

	Average Graduate Earnings	College Tier	2- vs 4-Year Degree
Panel A: Residual Theta Groups from Figure 3 and Table 4			
θ_{q45}, SAT_{q12}	11667.20***	4.34***	3.82***
	(395.40)	(0.19)	(0.42)
θ_{q12}, SAT_{q45}	6896.50***	2.39***	1.92***
	(359.96)	(0.11)	(0.17)
θ_{q45}, SAT_{q45}	27565.59***	18.88***	20.75***
	(189.32)	(0.56)	(1.45)
N	27333	28635	5523
θ_{q45}, SAT_{q12} = θ_{q12}, SAT_{q45}	0.000	0.000	0.000
θ_{q45}, SAT_{q12} = θ_{q45}, SAT_{q45}	0.000	0.000	0.000

Note: *Outcomes* - Earnings data are from the Chetty College Report Card and represent median child earnings in 2014 for the 1980-1982 birth cohorts, including non-college attendees. College Tier is an ordinal variable from 1-5 representing highly selective, selective, non-selective, two-year, and no college. 2- vs 4-Year College Degree is an ordinal variable indicating whether a student completed a 4-year degree, 2-year degree, No degree but some college, or no college. Degree completion is restricted to the 2017 graduating cohort. For both ordinal outcomes, coefficients are reported as odds ratios. *Models* - In Panel A, we use students identified from Table 4, with the bottom quintile Theta and SAT students as the reference group. The p-values report whether the high Theta/low SAT group is different from the low Theta/high SAT and high Theta/high SAT groups, respectively. Robust standard errors are shown in parentheses.

Table C1. Relationship between Grade Inflation and $\hat{\boldsymbol{\theta}}$

	[1]	[2]	[3]	[4]	[5]	[6]
Inflation (Std.)	0.362***	0.364***	0.353***	0.348***	0.392***	0.384***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Mean SAT (Std.)		0.718***		0.735***		0.694***
		(0.001)		(0.001)		(0.001)
N	1076252	1076252	1076252	1076252	1073812	1073812
Inflation Semi-Partial Correlation	0.187	0.139	0.182	0.137	0.218	0.166
SAT Semi-Partial Correlation		0.432		0.387		0.321
Grad Year FE	X	X	X	X	X	X
School FE			X	X	X	X
School*SCED FE					X	X

Note: All outcomes are standardized Theta scores. Models [1]-[2] include graduate year fixed effects, models [3]-[4] add school fixed effects, and models [5]-[6] include school-by-SCED course fixed effects (e.g., pre-calculus courses in specific schools). Robust standard errors are shown in parentheses. * 0.1 ** 0.05 *** 0.01

Appendix D: Simulation Study Details

This appendix provides the details of the three simulations studies used to explore the performance of our IRT-based approach under non-random selection. As noted in the main body, to evaluate how student course selection patterns might affect our $\hat{\theta}$ estimates, we conduct three simulation studies varying in their data generating processes. The simulation generated item responses for 1,000 students across 200 courses (items) with four ordinal response categories (0-3, corresponding to grades D through A). True student ability (θ) was drawn from a normal distribution $\theta \sim N(0, 2)$, chosen to approximate the empirically observed ability distribution in our real data ($\theta \sim N(0, 2.7)$).

For each course, three threshold parameters were generated by first drawing random values from U(0.5,1.5) and taking their cumulative sum. These thresholds were then centered by subtracting their mean and shifted by a random uniform draw from U(-3,1) to create variation in overall course difficulty. The final thresholds were multiplied by -1 to maintain the convention that higher θ values correspond to higher probability of better grades.

Course difficulties were calculated as the mean of their three threshold parameters, and step parameters were derived as the difference between each threshold and the course's mean difficulty. Response probabilities were then generated following the Partial Credit Model, where the log-odds of achieving each successive grade level is determined by the difference between student ability and the sum of the course difficulty and relevant step parameters. Random uniform draws were used to convert these probabilities into discrete grade responses.

To simulate realistic course-taking patterns, we retained only 28 courses per student, selected based on the proximity between student ability and course difficulty. Specifically, for each student, we identified the 1.5 × 28 courses with difficulty levels closest to their ability level, then randomly selected 28 of these courses to create the final response matrix – a probabilistic but strong selection mechanism.

The studies vary in the distribution of course difficulty relative to $\hat{\theta}$ and the mechanism by which courses are selected at the student level. In all cases, the average course difficulty is lower than the average $\hat{\theta}$, reflecting the distributions found when we fit the model to real data. These simulations are intended to outline the extent to which key aspects of the course selection process influence $\hat{\theta}$, but are simplified relative to our real data analysis in ways that facilitate faster model estimation, such as the smaller sample size and items being on a four-point scale instead of five. Given the non-trivial amount of time needed to run each replication of each study, we run 50 replications per condition.

Item difficulty distributions and thresholds

In study 1, the item difficulty distribution is U(-5, 5), representing an effectively unrestricted difficulty range in which all students can select into courses close to their θ. In study 2 and 3, it is U(-3, 1), meaning that the courses are on average slightly easy compared to the mean θ, and that there is range restriction for very high and low θ students selecting into courses adjacent to their

θ. In all cases, courses are scored on a 0-3 scale. The thresholds are spaced from one another via random magnitudes drawn from the $U(0.5, 1.5)$ distribution.

Course selection mechanism

In all three studies, each student is simulated as having taken 28 courses of the 200 total. Scores in all courses are simulated according to each examinee's θ and the item's threshold parameters, after which selection is applied by which 172 of the scores are converted to missing values. The selection mechanisms are as follows.

For studies 1 and 2, a "core" of 8 courses is chosen at random from the 10th through 60th percentile of course difficulty. Every student is then selected into five of these eight courses at random. The remaining 23 courses for each student are the 23 courses whose difficulty is closest (based on absolute difference) to that student's θ (representing an extremely strong/deterministic selection mechanism). For study 3, the first step is skipped – there is no core. Each student just takes the 28 courses whose difficulty is closest to their θ. All scenarios therefore represent a level of selection that is as extreme as possible.

Evaluation

For each study, we compare the extent to which (a) $\hat{\theta}$ based on the 28 non-randomly selected courses recovers $\hat{\theta}$ based on all 200 courses to (b) the extent to which GPA from the 28 courses recovers GPA from all 200. For each study, we compute and report several comparison statistics, as follows.

Bias

Bias is the mean signed difference between an estimate of a "gold standard" value and that gold standard value. It tells us the extent to which the estimate is systematically too high or too low across the entire distribution of observed values. For a replication *i* of a condition *c*, bias for θ estimates is computed as:

$$bias_{\theta ci} = \frac{\sum_{p=1}^{n} \hat{\theta}_{p,obs} - \hat{\theta}_{p,full}}{n} \tag{B1}$$

where p indexes students, *n* is the number of students in the replication (in this case 1000), $\hat{\theta}_{p,obs}$ is student p's estimated θ based on 28 selected courses, and $\hat{\theta}_{p,full}$ is their estimated θ based on all 200 courses. Analogously, bias for GPA is computed as follows:

$$bias_{GPAci} = \frac{\sum_{p=1}^{n} GPA_{p,obs} - GPA_{p,full}}{n} \tag{B2}$$

We report the mean of these two measures of bias for the 50 replications of each simulation condition.

Root-mean-square error

Root-mean-square error (RMSE) is the square root of the average squared difference between an estimate and its gold standard. It tells us how far off an estimate is from its gold standard in absolute terms across the entire distribution of observed values. Carrying forward the subscripts from equations B1 and B2, RMSE for θ and GPA are:

$$RMSE_{\theta ci} = \sqrt{\frac{\sum_{p=1}^{n}\left(\hat{\theta}_{p,obs} - \hat{\theta}_{p,full}\right)^2}{n}} \tag{B3}$$

$$RMSE_{GPAci} = \sqrt{\frac{\sum_{p=1}^{n}\left(GPA_{p,obs} - GPA_{p,full}\right)^2}{n}} \tag{B4}$$

Like bias, we report the mean RMSE for both θ and GPA for the 50 replications of each condition.

Normalized root-mean-square error

Normalized RMSE (NRMSE) divides the RMSE by the standard deviation of the observations, facilitating comparison of the RMSEs for θ and GPA that are otherwise on different scales. This is also reported as a mean across 50 replications for each condition.

Concordance correlation coefficient

The concordance correlation coefficient (CCC) is a measure of agreement between two variables that accounts for both their Pearson correlation and the differences in the first and second moments of their distributions. It accounts for both precision and accuracy in comparing the variables. We report the CCC for $\hat{\theta}$ based on the 28 non-randomly selected courses with $\hat{\theta}$ based on all 200 courses, and the CCC for GPA based on selected and full courses as well. We report the means of these two CCCs across all 50 replications of each condition.

Correlations between observed scores

We also report mean Pearson correlations between the "observed" $\hat{\theta}$ and GPA (i.e. based on 28 courses), and the standard deviation of these correlations, across the 50 replications of each condition.

Recovery of data-generating θ

Finally, we report the extent to which the standardized "full" (all 200 courses) $\hat{\theta}$ and GPA recover the standardized data-generating θ. Here, we report the NRMSE, but instead of comparing an estimate from 28 courses to an estimate from 200, we compared the estimate from 200 to the true data-generating θ. We produce this comparison for both $\hat{\theta}$ based on 200 courses and GPA based on 200 courses, with all variables standardized. This represents the extent to which these two approaches to scoring students' transcripts differ in their representation of the underlying data-generating measure of transcript quality, free of selection issues.