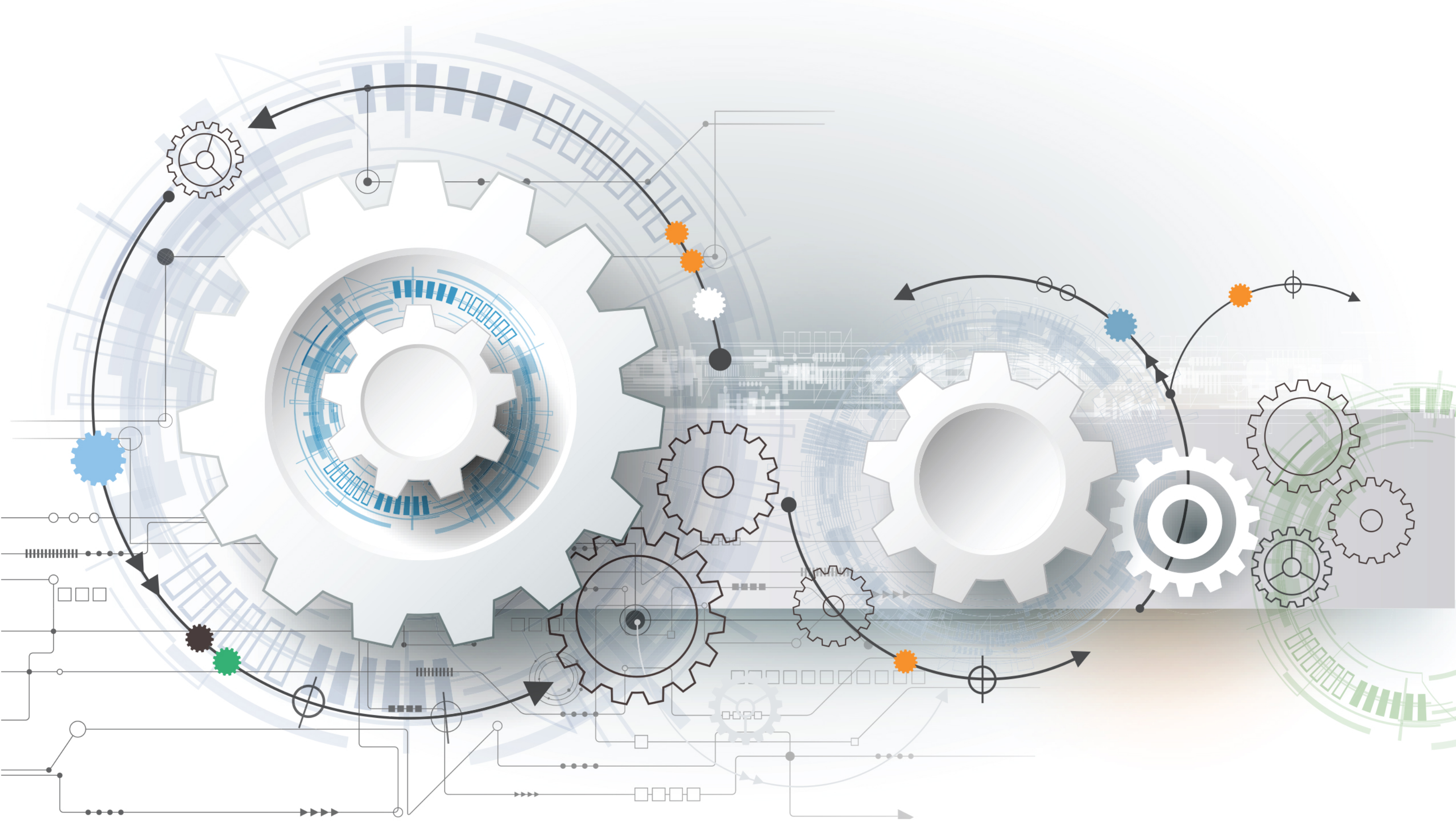
Working Paper
2018-4

Examination of Indices of High School Coursework and Grades Based on the Graded Response Model

JEFF ALLEN, PHD
KRISTA MATTERN, PHD



ACT.org

ABOUT THE AUTHORS

Jeff Allen is a statistician in the Research division at ACT. He specializes in longitudinal research linking test scores to educational outcomes and student growth models.

Krista Mattern is a senior director in Validity and Efficacy Research specializing in the validity and fairness of assessment scores as well as more general issues in higher education such as enrollment, persistence, and graduation

ACT WORKING PAPER SERIES

ACT working papers document preliminary research. The papers are intended to promote discussion and feedback before formal publication. The research does not necessarily reflect the views of ACT.

R1706

Abstract

We examined indices of high school coursework and grades based on the graded response model (GRM). The indices varied by inclusion of ACT® test scores and whether high school courses were constrained to have the same difficulty and discrimination across groups of schools. The indices were examined with respect to skewness, incremental prediction of college degree attainment, and differences across racial/ethnic and socioeconomic subgroups. The most difficult high school courses to earn an “A” grade included calculus, chemistry, trigonometry, other advanced math, physics, algebra 2, and geometry. The GRM-based indices were less skewed than simple HSGPA and had higher correlations with ACT Composite score. The index that included ACT test scores and allowed item parameters to vary by school group was most predictive of college degree attainment but had larger subgroup differences. Implications for implementing multiple measure models for college readiness are discussed.

Keywords: graded response model, high school GPA, college readiness, college degree attainment, academic rigor

1. Introduction

The simple high school grade point average (HSGPA) is a popular measure of college readiness and is among the most important factors influencing college admissions decisions (Clinedinst & Koranteng, 2017). While research has established that HSGPA is predictive of first-year college GPA (FYGPA), it is generally understood that the level and intensity of courses taken are also important factors for understanding college readiness (Adelman, 1999; Adelman, 2006). For admissions and placement decisions, colleges will often employ weighting of course grades and/or award bonus points for advanced courses (Sadler & Tai, 2007). Use of weights or bonus points also encourages students to take challenging courses in high school (Klopfenstein & Lively, 2016).

The choice of how to weight high school courses and award bonus points has received attention from researchers. Three families of approaches include: (1) rational; (2) empirical, using predictive linkages to outcomes such as college grades; and (3) empirical, using scaling methods based on Item Response Theory (IRT). Each family of approaches includes models that account for between-school differences in grading practices or the intensity or quality of courses. However, for practical (e.g., data limitations) or policy reasons, between-school differences are not always accounted for.

Rational approaches are driven by policy or judgment. For example, some high schools use a weighted GPA scale that awards one extra point for advanced courses, such as those designated as Advanced Placement (AP; PrepScholar, 2018). Students who earn a B in an AP course are awarded the same points as a student who earns an A in the same course without the AP designation. Rational approaches can be evaluated by the empirically-based methods (c.f., Hansen, Sadler, & Sonnert, 2016) and through policy analysis (c.f., Klopfenstein & Lively, 2016).

Assignment of weights and bonus points can also be based on empirical relationships of coursework and grades to outcomes, such as FYGPA. For example, the academic rigor index (ARI) was developed by relating high school coursework indicators to FYGPA (Wyatt, Wiley, Camara, & Proestler, 2011; Beatty, Sackett, Kuncel, Kiger, Rigdon, Shen, & Walmsley, 2012). More recently, an index of high school academic rigor was developed by optimizing the prediction of FYGPA based on high school courses taken, grades, and indicators of advanced coursework (HSAR index; Allen, Ndum, & Mattern, 2017). Indices that optimize the prediction of FYGPA may not perform as well for other outcomes such as postsecondary degree completion. Hierarchical modeling (students nested within high schools) can be used to develop indices with school-specific effects of coursework and grades. Neither the ARI nor the HSAR index accommodates school-specific effects.

Another family of approaches for weighting course grades and awarding bonus points uses scaling methods based on Item Response Theory (IRT). The graded response model (GRM; Samejima, 1969) has been used to obtain an alternative weighting of HSGPA (Hansen, Sadler, & Sonnert, 2016). This model allows difficulty to vary across high school courses, allows for the difference between letter grades to vary for each course (e.g., the difference between A and B can be different than the difference between B and C), and allows the reliability of grades to vary across courses. The scaling models can account for between-school differences in course difficulty. For example, Bassiri and Schulz (2003) used ACT test scores as common items across high schools to create a high school course difficulty scale using the Rasch rating scale model (Andrich, 1978). Item parameters for course grades were allowed to vary across schools.

The two families of empirically-based approaches (outcome prediction and scaling) have different data requirements. To develop an outcome-based index, one must link the high school transcript data to relevant outcome data, such as FYGPA or other college outcomes. Once the prediction model is established, the model can be applied to students with “predictors” (e.g., course grades or coursework indicators) derived from their high school transcript data. The predictor data must be non-missing or otherwise imputed. In contrast, the scaling-based approaches require no links to relevant outcome data. Further, the scaling-based approaches are more flexible in how high school transcript data are treated. Indices can be estimated using the available transcript data, and missing data only causes a loss of precision. Scaling-based approaches accommodate missing data in a manner analogous to students taking exams of different lengths.

In this study, we produce different GRM-based indices of high school coursework and grades. We evaluate the indices with respect to (1) skewness, (2) prediction of postsecondary degree attainment, and (3) differences across racial/ethnic and socioeconomic subgroups. The results for the GRM-based indices are contrasted to those for HSGPA and the HSAR index.

2. Methods

2.1 Sample

The sample includes students who took the ACT test in 11th or 12th grade, were projected to complete high school in 2010, and attended a high school in the United States (n=1,517,656). To ensure that measures of school mean achievement are not based only on students who elect to take a college admissions test, students must have attended a public high school where at least 90% of students took the ACT test (n=185,386). High school coursework and grades data, demographics, and educational plans are collected when students register to take the ACT test. To ensure that the measures of high school coursework and grades are based on adequate data, students must have provided grades for at least 15 of 30 courses (n=50,058) to be included in the analysis sample. Because the students who provided grades data are generally higher-achieving and different on socio-demographic variables, propensity score weights (Rosenbaum, 1987) were applied to the analysis sample to make it more representative of the population of all high school students.[1]

The students represented 1,030 high schools from 32 states. The weighted sample includes students from the Midwest (56%), South (29%), West (15%), and Northeast (<1%) regions. The weighted sample is 52% female, 48% male, 63% White, 17% African American, 7% Hispanic, 3% Asian, 4% other race/ethnicity, and 6% missing race/ethnicity. Most students in the sample expected to complete a bachelor's degree (40%), one or two years of graduate study (17%), or a doctorate or professional degree (30%). The remainder expected to complete an associate's degree (5%), certificate program (1%), other type of degree (2%), or did not respond to the questionnaire item (4%).

2.2 High school grades, coursework, and ACT test scores

For 30 different high school courses, students are asked to report the grade they earned in each course already taken, with five options (A, B, C, D, or F). HSGPA was determined by averaging grades reported by students across the 30 high school courses. When students register for the ACT test, they are also asked whether they have taken advanced placement, accelerated, or honors courses in English, mathematics, social studies, natural sciences, or foreign languages. Binary indicators for each type of advanced coursework were used. As described later, the course

[1] Propensity scores were based on a logistic regression of inclusion in the analysis sample, with the following covariates: ACT Composite score, HSGPA, gender, race/ethnicity, family income, school mean ACT score, school percent eligible for free or reduced lunch, and the degree attainment outcomes. The weights were set as the inverse of the propensity score.

grades and indicators for advanced coursework are used to derive the GRM-based indices of high school coursework and grades.

The ACT test is designed to measure academic skills necessary for education and work after high school, and the content of the tests is related to major curriculum areas (ACT, 2014). The ACT includes English, mathematics, reading, science, and an optional writing test. The tests focus on knowledge and skills attained as the cumulative effect of school experience. The tests are oriented towards the general content areas of college and high school instructional programs. The ACT Composite score is the average of the four ACT subject area scores from the multiple choice portion of the test (English, mathematics, reading, and science), and each of these scores is reported on a 1-36 scale. As described later, ACT test scores are used to derive some of the GRM-based indices. ACT Composite score is used to examine predictors of college degree attainment and subgroup differences of college readiness measures.

Students also provided their gender, race/ethnicity, and family income level when registering for the ACT test. Each high school's mean achievement was estimated by their mean ACT Composite score (prior to removing students who did not provide adequate coursework information). These variables are used for subgroup analyses described later.

2.3 The graded response model

The GRM is a polytomous IRT model appropriate for ordered categorical responses, such as course grades (e.g., A, B, C, D, F) and Likert scale responses (e.g., Strongly Disagree, Disagree, etc.) (Samejima, 1997). Let Y_{ij} represent an ordered categorical response for examinee j to item i with K_i response options labeled 0, 1, 2, .. K_i-1. The probability distribution of Y_{ij} is modeled as a function of examinee ability (θ_j) and item parameters $\alpha_i, \beta_{1j}, \beta_{2j}, .. \beta_{(K_i-1)j}$. For k=1, 2, … K_i-1:

$$\Pr\left(Y_{ij} \geq k \mid \theta_j, \alpha_i, \beta_{1j}, \beta_{2j}, ..\beta_{(K_i-1)j}\right) = \left(1 + \exp\{-\alpha_i(\theta_j - \beta_{kj})\}\right)^{-1}$$

As an example, consider grades in a hypothetical high school course for students of average ability ($\theta_j = 0$). Assume model parameters of α=1.5, β_1=-2, β_2=0, and β_3=1. Table 1 provides the probabilities of earning each grade level when $\theta_j = 0$. Among students with average ability, 18.2% earn an A, 31.8% earn a B, 45.2% earn a C, and 4.7% earn a D or F. Figure 1 plots the probabilities of each course grade, across ability levels from -3 to 3.

Table 1. Demonstrating the GRM for High School Grades

Grade	Response label (k)	Difficulty (β_k)	$\Pr(Y \geq k)$	$\Pr(Y = k)$
D/F	0		1.000	0.047
C	1	-2.0	0.953	0.453
B	2	0.0	0.500	0.318
A	3	1.0	0.182	0.182

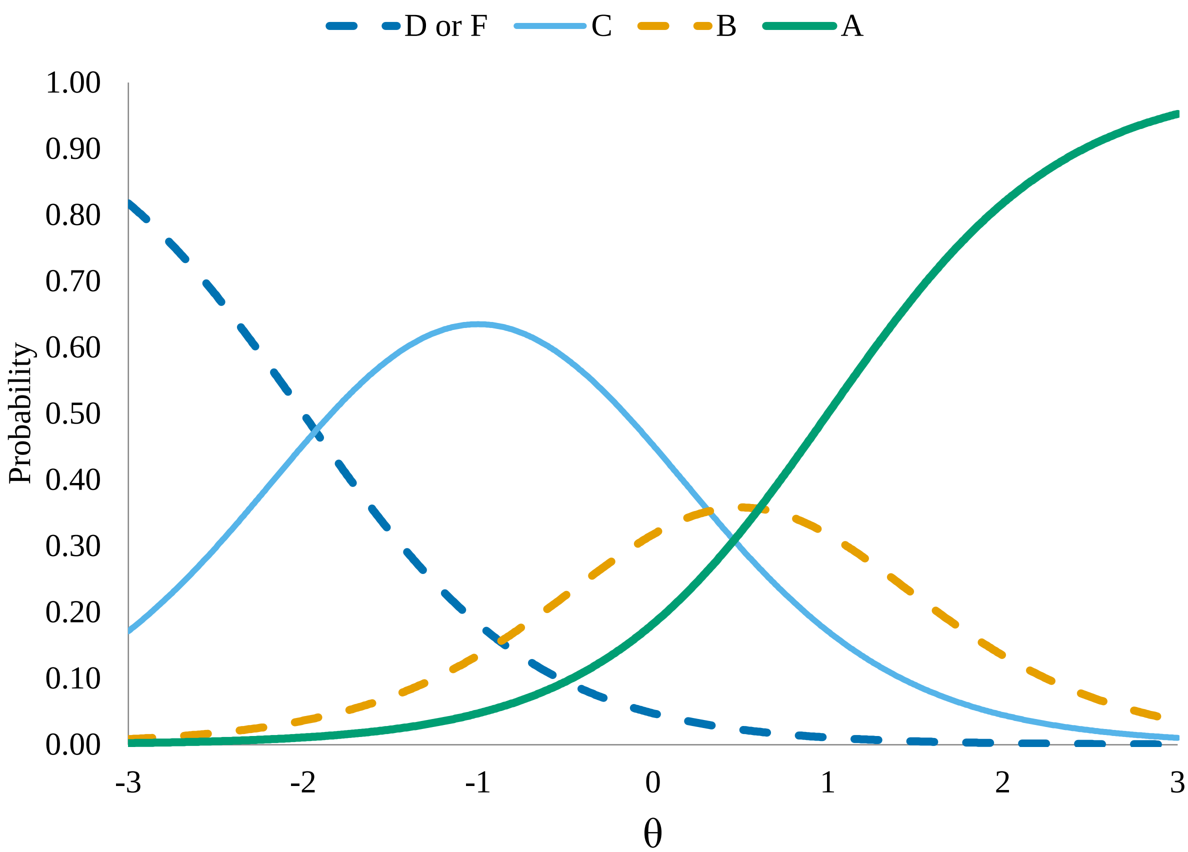


Figure 1. Hypothetical GRM-based probabilities of course grades, by ability

Latent ability for examinee *j* (θ_j) represents "a hypothetical construct underlying certain behavior" (Samejima, 1997), in this case, behavior that drives course-taking and performance. It is assumed to have population mean 0 and standard deviation 1. While it's likely an oversimplification to assume that course-taking and performance behavior is one-dimensional, we treat it as such to be consistent with the operationalization of HSGPA and to derive a single measure of coursework and grades. The α parameter measures the rate at which response probabilities change with ability and is also referred to as discrimination (Samejima, 1997). The β parameters measure the difficulty of obtaining each grade. In Figure 2, we plot the original probability of earning an A in the hypothetical course, as well as the probability curves corresponding to unit shifts in the α and β_3 parameters. Increasing α results in a steeper probability curve. Increasing β results in a curve that is shifted to the right, indicating greater difficulty. For the original curve, θ=1 is needed to have a 0.50 probability of earning an A. With β_3 shifted one unit to the right, θ=2 is needed to have a 0.50 probability of earning an A.

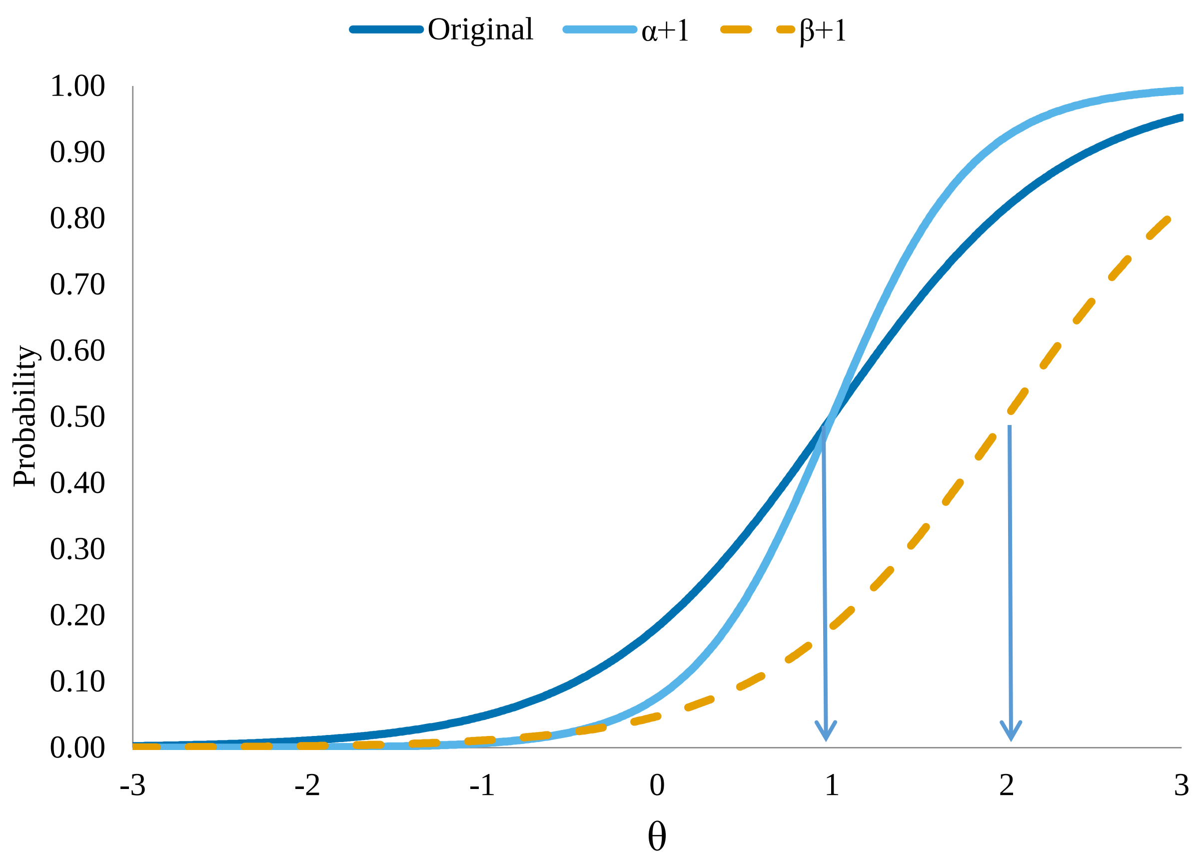


Figure 2. Hypothetical probability of “A”

2.4 GRM-based indices

The GRM can be fit to the high school coursework and grades data collected on the ACT registration form. Modeling decisions include: (1) which course grades to include, (2) inclusion of advanced coursework indicators, (3) if and how to include ACT test scores, and (4) if and how to let model parameters vary across high schools.

For each GRM, we chose to use course grades for all 30 courses collected on the ACT registration form. Table 2 lists the 30 courses and provides course grade response rates and frequency distributions. Very few students reported F grades, so grades D and F were combined into one category. We also used the five binary advanced coursework indicators. Thus, ability estimates are affected by students' grades in the mix of courses for which they reported course grades, as well as whether they reported taking one or more advanced courses in English, math, social studies, natural science, or foreign language.

ACT test scores can be grouped into ordered categories and used in the GRM. Inclusion of standardized measures like ACT test scores has the potential to illuminate differences in course difficulty across high schools. ACT test scores can serve as "common items" while parameters for coursework and grades can vary across high schools or groups of high schools. Inclusion of ACT test scores also has potential to help calibrate parameters for coursework and grades, even if the parameters are constrained to be the same across all schools. A downside of including ACT test scores is that the ability estimates are then driven both by coursework and test scores, calling the dimensionality assumption into greater question. Further, for some uses (e.g., multiple measure models for college admissions, research targeting specific coursework outcomes), keeping measures of high school transcript data distinct from test scores is desirable. We fit GRMs (1) without ACT test scores, (2) with ACT scores used for calibration and ability estimates, and (3) with ACT scores used only for calibration.

Table 2. Course Grade Response Rates and Distributions

Course	Response rate	Grade distribution			
		A	B	C	D/F
English 9	0.998	0.385	0.380	0.187	0.047
English 10	0.997	0.362	0.403	0.188	0.047
English 11	0.985	0.357	0.389	0.197	0.057
English 12	0.391	0.372	0.367	0.198	0.063
Other English	0.157	0.598	0.263	0.108	0.030
Algebra 1	0.992	0.400	0.313	0.216	0.072
Geometry	0.979	0.307	0.356	0.249	0.088
Algebra 2	0.953	0.308	0.354	0.249	0.090
Trigonometry	0.427	0.369	0.364	0.196	0.071
Calculus	0.088	0.459	0.333	0.161	0.047
Other math beyond Algebra 2	0.362	0.384	0.364	0.192	0.061
Computer Math/ Science	0.145	0.567	0.270	0.124	0.038
Physical, Earth, General Science	0.819	0.420	0.350	0.185	0.045
Biology	0.987	0.344	0.393	0.205	0.058
Chemistry	0.873	0.295	0.355	0.255	0.095
Physics	0.414	0.336	0.364	0.216	0.084
U.S., American History	0.984	0.427	0.356	0.173	0.044
World History, Civilization	0.857	0.418	0.363	0.174	0.045
Other History	0.271	0.428	0.362	0.172	0.038
Government, Civics, Citizenship	0.716	0.425	0.346	0.186	0.043
Economics, Consumer Econ.	0.561	0.422	0.340	0.184	0.053
Geography	0.518	0.467	0.328	0.162	0.043
Psychology	0.280	0.476	0.327	0.146	0.051
Spanish	0.581	0.409	0.363	0.182	0.047
French	0.132	0.436	0.333	0.175	0.055
German	0.045	0.455	0.323	0.165	0.057
Other Language	0.055	0.558	0.276	0.130	0.035
Art	0.432	0.714	0.206	0.065	0.015
Music	0.320	0.835	0.117	0.040	0.009
Drama/Theater	0.145	0.754	0.177	0.055	0.014
High school advanced coursework	Response rate	Yes	No		
English	0.590	0.679	0.321		
Mathematics	0.551	0.645	0.355		
Social Studies	0.546	0.613	0.387		
Natural Sciences	0.544	0.611	0.389		
Foreign Languages	0.441	0.422	0.578		

Course difficulty is expected to vary by school and instructor. Our data set does not include course instructor (e.g., high school teacher), but does include school. Therefore, with sufficient within-school sample size, we could fit GRMs specific to schools (e.g., each course within each school is treated as a distinct item). Alternatively, simpler versions of the GRM can be fit for groups of schools. By grouping schools, we ensure sufficient sample size and schools can be grouped in a manner that facilitates exploration of systematic differences in course difficulty and discrimination. We assigned each school to one of three groups based on mean ACT Composite score: (1) lower achieving, (2) middle achieving, and (3) higher achieving. Prior to deleting students who provided fewer than 15 course grades, the group boundaries were set to result in an equal number of students per group. After deleting students with fewer than 15 course grades, there are relatively more students in the higher achieving school group because nonresponse is related to school mean achievement. The sample sizes per group are: 12,650 for lower-achieving schools; 15,596 for middle-achieving schools; and 21,812 for higher-achieving schools.

We fit pooled GRMs (with coursework and grades parameters constant across schools) and group-specific GRMs (with coursework and grades parameters specific to high school group). With three variations of ACT score use and two variations for treatment of school groups, there are six total GRM-based indices calculated as ability estimates and denoted θ_1, θ_2,…, θ_6 (Table 3). The GRM models were fit using the `grm` function of the R package `ltm` (Rizopoulos, 2006). GRM-based indices (ability estimates) were obtained using the `factor.scores` function. To obtain the ability estimates that only used ACT scores for calibtation (θ_3 and θ_6), the `factor.scores` function was applied to a data set with the ACT score variables set to missing.

Table 3. College Readiness Measures Included in Analysis

Measure	Description
HSGPA	Unweighted average across 30 high school courses. Based on student-reported grades.
ACT Composite Score	Standardized test score summarizing performance in English, math, reading, and science on the ACT test.
HSAR index	Index derived from multiple linear regression prediction of FYGPA (Allen et al., 2017). Based on 30 course outcomes and 5 indicators of advanced coursework.
GRM HSGPA-1 (θ_1)	Latent ability estimated from GRM, where 30 course grades and 5 indicators of advanced coursework are treated as items. Items are common across all high schools.
GRM HSGPA-2 (θ_2)	Same as GRM HSGPA-1, but grouped ACT test scores are also included as items.
GRM HSGPA-3 (θ_3)	Same as GRM HSGPA-2, but grouped ACT test scores are only used for calibrating the GRM coursework parameters, not for estimating latent ability.
GRM HSGPA-4 (θ_4)	Latent ability estimated from GRM, where 30 course grades and 5 indicators of advanced coursework are treated as items. Items are specific to high school group, where group is determined by aggregate performance on the ACT test.
GRM HSGPA-5 (θ_5)	Same as GRM HSGPA-4, but grouped ACT test scores are also included as items.
GRM HSGPA-6 (θ_6)	Same as GRM HSGPA-5, but grouped ACT test scores are only used for calibrating the GRM coursework parameters, not for estimating latent ability.

2.5 The high school academic rigor (HSAR) index

The HSAR index is an empirically-based predictor of FYGPA (Allen et al., 2017). Using a nominal parameterization of high school course outcomes, the HSAR index capitalizes on differential contributions across courses and nonlinear relationships between course grades and

FYGPA. Most of the inputs to the HSAR index are the same as those used for the GRM-based indices (high school course grades and indictors for advanced coursework), but coursework variable also include "not taken" as a response category. Relative to HSGPA and ACT Composite score, the HSAR index was the strongest predictor of FYGPA, but it only led to a modest incremental prediction of FYGPA over a base model with HSGPA and ACT Composite score (Allen et al., 2017).

The HSAR index was calculated for students in the current study sample using the scoring parameters estimated previously using over 109,000 students who completed high school between 2006 and 2015 (Allen, Ndum, & Mattern, 2018). Along with HSGPA, ACT Composite score, and the GRM-based indices, the HSAR index is used to examine predictors of college degree attainment and subgroup differences of college readiness measures.

2.6 Degree attainment data

For students in the sample, degree attainment records were obtained through the National Student Clearinghouse (NSC, 2018). The NSC's participating institutions enroll over 98% of all students in public and private institutions in the US, and the degree verification service represents nearly 94% of US four-year degrees (NSC, 2018). Degrees were tracked from fall 2010 through summer 2017. Three graduation outcomes were defined as: (1) any degree or certificate, (2) bachelor's degree or higher, and (3) post-bachelor's degree.

2.7 Analyses

Analyses were conducted to compare the measures of high school coursework and grades (HSGPA, HSAR index, and the GRM-based measures) on (1) skewness of frequency distribution, (2) incremental prediction of college degree attainment, and (3) differences across racial/ethnic and socioeconomic subgroups. For each measure, skewness is calculated and histograms are presented to compare the shapes of the distributions.

Multilevel logistic regression is used to examine incremental prediction of the three college degree attainment outcomes. The first model includes HSGPA and ACT Composite score as baseline predictors. Student nesting within high schools is modeled with random intercepts. The incremental contribution of each alternative measure (HSAR index, each GRM-based measure) is then tested by including each in subsequent models. For each model, the overall odds ratio (*OOR*; Allen & Le, 2008) is used to describe overall effect size, and *OOR* values are compared to the baseline model. Models are also fit to examine how different GRM measures explain variation in outcomes that would otherwise be explained by background variables (gender, race/ethnicity, family income, school mean ACT Composite score, and school percent eligible for free or reduced lunch).

To examine differences in measures (HSGPA, HSAR index, and each GRM-based measure) across racial/ethnic and socioeconomic subgroups, we calculated the standardized difference in mean scores (*d*) for White versus African American and Hispanic students. Correlations with family income and school mean ACT Composite score are also presented for each measure.

To deal with intermittently missing values of family income and race/ethnicity, multiple imputation is used to generate 5 complete data sets using the R package `MICE` (van Buuren & Groothuis-Oudshoorn, 2011). The SAS PROC MIANALYZE procedure (SAS Institute Inc., 2008) is used to combine the results across the multiple imputed data sets and the confidence intervals of the logistic regression estimates include variation within and between data sets.

3. Results

Different versions of the GRM were fit to the high school coursework and grades data, varying by inclusion of ACT test scores (no use of ACT score, inclusion as items for estimating ability, inclusion as items for calibration only) and whether schools were pooled or grouped by school mean achievement level. The results of the GRM model provide information about the difficulty of high school courses. Figure 3 summarizes results from the GRM model that grouped schools and used ACT scores as items (note that results are only shown for 10 of the 30 courses). Following the graphical approach used by Hansen et al. (2016), the figure shows the difficulty parameter (β) estimates, indicating the ability levels (θ) associated with a 50% chance of earning each grade or higher. As expected, difficulty parameters vary by course and are highest for the higher achieving schools. For example, earning a “B” in Calculus at a higher-achieving school is as difficult as earning an “A” in English 11 at a lower-achieving school.

Correlations, means, standard deviations, and skewness of the college readiness measures and the degree outcomes are provided in Appendix Table A1. Examining correlations with HSGPA and ACT Composite scores may help us understand which GRM-based measures of ability are more likely to provide unique information about college readiness. As expected, the GRM-based measures of ability are highly correlated with HSGPA, with correlations ranging from 0.911 for θ_5, which was based on model that grouped schools and included ACT scores, to 0.969 for θ_1, which was based on a model that pooled schools and did not include ACT scores. Conversely, correlations with ACT Composite score were highest for θ_5 (r=0.828) and lowest for θ_4 (r=0.579), which was based on a model that grouped schools and did not include ACT scores. Ability estimates that only used ACT scores for calibration (θ_3 and θ_6) have more moderate correlations with HSGPA and ACT Composite score.

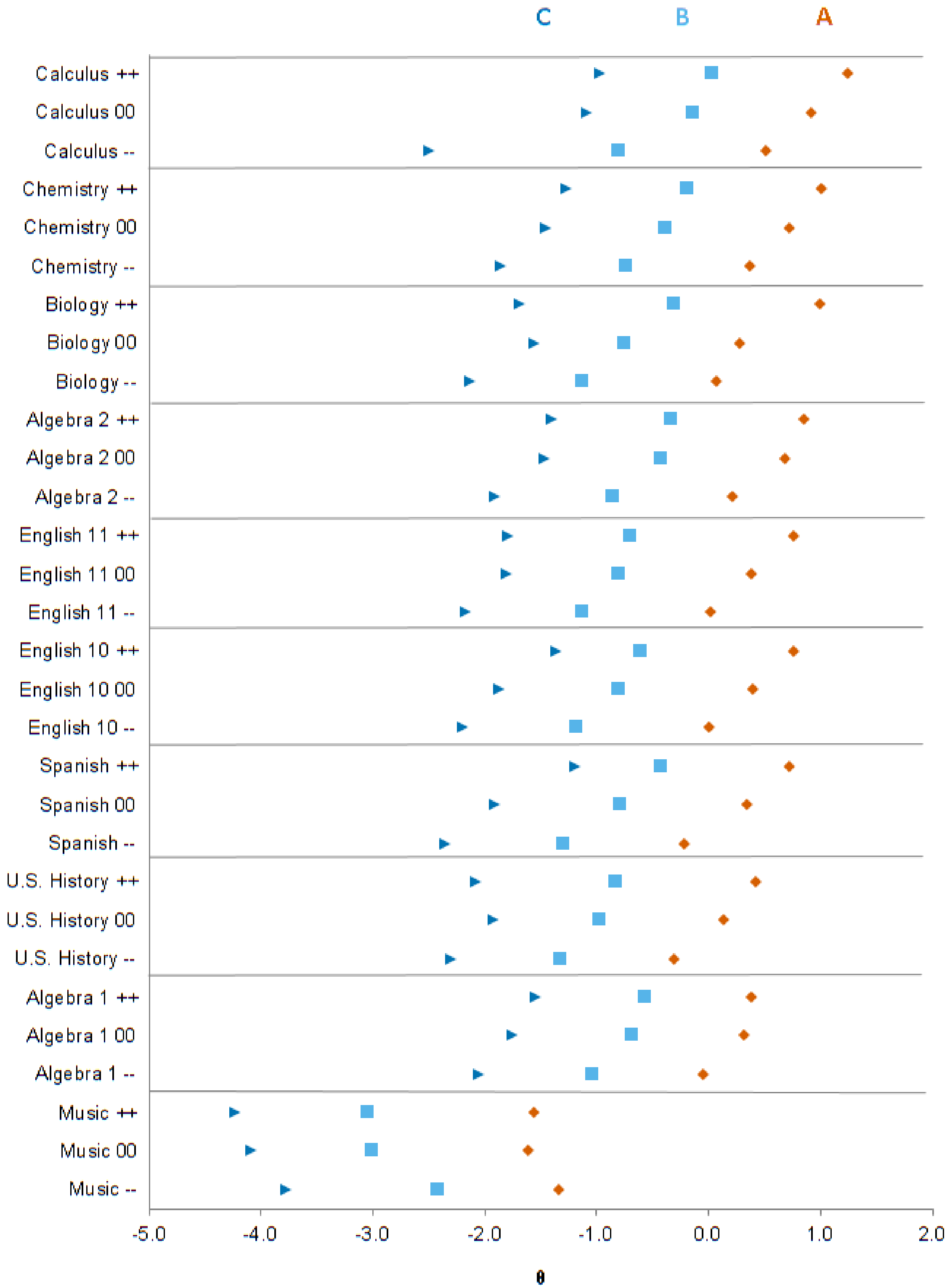


Figure 3. Ability (θ) needed for 0.50 probability of earning each grade or higher

Note: ++ =higher achieving schools, 00=middle achieving schools, --=lower achieving schools

About 47% of the weighted sample earned any postsecondary degree or certificate within 7 years after high school, while 38% earned a bachelor's degree or higher, and 4% earned a degree after a bachelor's. Across the college readiness measures, correlations with bachelor's (or higher) degree attainment ranged from 0.428 (for θ_4, grouped school model without ACT scores) to 0.530 (for θ_5, grouped school model with ACT scores). The correlations for the traditional measures of college readiness were 0.459 for HSGPA and 0.468 for ACT Composite score. The GRM-based measures outperformed the traditional measures for predicting degree attainment when ACT scores were included in the model, even if only for calibration.

HSGPA was negatively skewed (skewness=-0.57, Figure 4), with the peak of the distribution occurring at the maximum score, corresponding to HSGPA=4.0 and a z-score of about 1.4. College readiness measures that are heavily skewed, or have many "tied" observations, are less able to distinguish students for admissions and placement. Traditionally, this has been one of the reasons that ACT and SAT test scores have complemented high school grades as college readiness measures (Sawyer, 2010). The GRM-based measures from the pooled school models (θ_1, θ_2, θ_3) are less skewed and thus are more bell-shaped than HSGPA, with skewness ranging from 0.074 (θ_1) to 0.220 (θ_2) (Figure 5). The measures still have spikes in the distribution for students who reported all A's, but the spikes are much less pronounced relative to the spike for HSGPA. The GRM-based measures from the grouped school models (θ_4, θ_5, θ_6) are even less skewed and are more bell-shaped, with skewness ranging from 0.069 (θ_6) to 0.135 (θ_5) (Figure 6). The distributions for θ_4, θ_5, and θ_6 can be thought of as mixtures of distributions for the three school groups. The group differences are most pronounced for θ_5, which includes ACT scores as items (mean=-0.64 for group 1, mean=0.04 for group 2, and mean=0.51 for group 3).

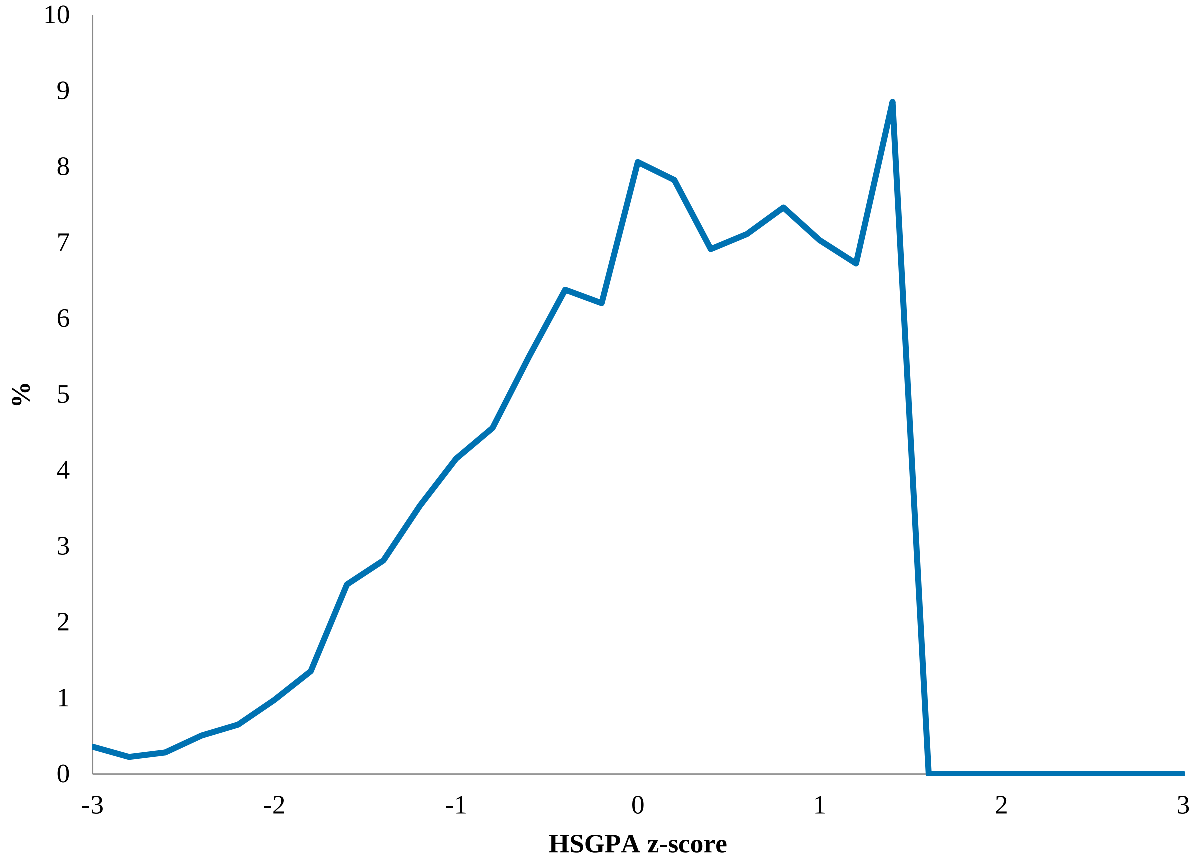


Figure 4. HSGPA distribution

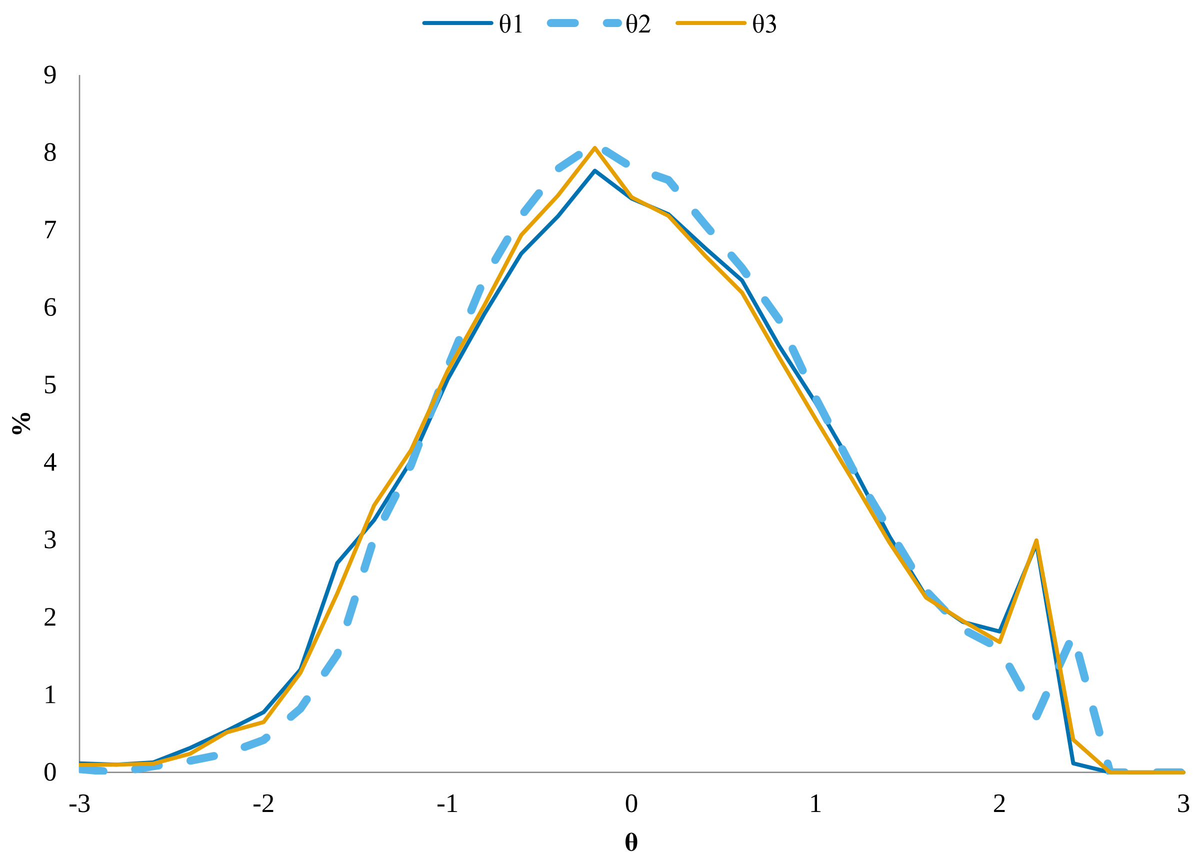


Figure 5. Distributions of GRM-based ability estimates from pooled school models (θ_1, θ_2, θ_3)

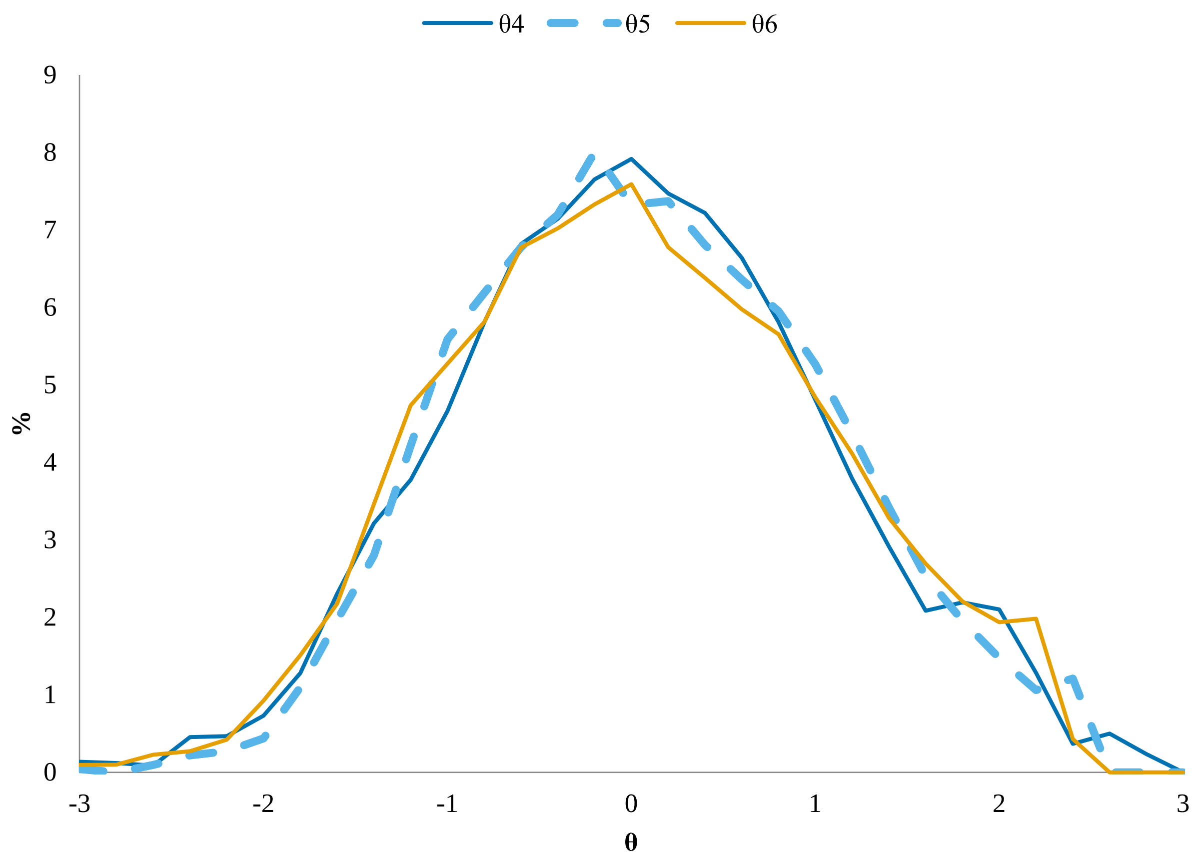


Figure 6. Distributions of GRM-based ability estimates from grouped school models (θ_4, θ_5, θ_6)

Predictors of earning a bachelor's degree or higher within seven years after high school were examined (Appendix Table A2). The base model (model 1) included the traditional predictors (HSGPA and ACT Composite score), and subsequent models examined the incremental prediction of the HSAR index (model 2) and the GRM-based measures (models 3-8). For the base model, the overall odds ratio (*OOR*) was 3.92. This means that the odds of earning a bachelor's degree or higher increase by a factor of 3.92 for each standard deviation increase in the model's linear predictor (e.g., weighted combination of HSGPA and ACT Composite score). For example, suppose the probability of earning a degree is 0.50 for a student with average HSGPA and ACT Composite scores. This corresponds to odds of 1.0. A student whose linear predictor is one standard deviation above the mean would have odds of earning a degree of 3.92, which corresponds to a probability of 0.80.[2]

Each of the alternative college readiness measures (HSAR index and each GRM-based index) led to a modest increase in the *OOR*. The largest *OOR* (4.15, model 7) was obtained with θ_5, which is based on the grouped school model with inclusion of ACT scores. In model 7, θ_5 is the strongest predictor of degree attainment (adjusted OR=3.14), and HSGPA and ACT Composite scores have such smaller effects due to high overlap with θ_5, which is a function of course grades and ACT scores. Model 8 tests the incremental prediction of θ_6, which is based on the grouped school model with ACT score calibration. The *OOR* (4.12) is slightly smaller than the *OOR* for model 7, and the relative contribution of ACT Composite score (adjusted OR=1.48) is stronger because ACT scores are not also included in the derivation of θ_6. The use of θ_6 leads to a 5% increase in predictive strength over the base model. Both θ_5 and θ_6 outperformed the HSAR index, showing that the GRM model can generate ability estimates that predict better than indices designed to optimally predict other outcomes (e.g., FYGPA).

Using ACT scores for calibration enhances the predictive strength of GRM-based indices when the model is grouped by school (compare results for θ_6 to θ_4), but not when the model is pooled across schools (compare results for θ_3 to θ_1). This suggests that calibrating high school course grades using standardized test scores has the greatest potential when the GRM model parameters are school-specific. GRM-based indices are most predictive when ACT scores are included as items, but such indices may not be desired because they mix grades with test scores. θ_4 was based on the grouped-school model without ACT scores (as items or for calibration) and was least useful for predicting bachelor's degree attainment. *OOR* results for the other degree

---

[2] odds = *p*/(1-*p*) and *p* = odds / (1+odds).

outcomes (any degree and post-bachelor's degree) are also provided in Table A2, and the pattern of results is very similar to what is observed for the bachelor's or higher outcome.

Relative to HSGPA, the alternative measures of college readiness had mostly comparable racial/ethnic differences and similar correlations with family income and school mean ACT score (Table 4). The GRM-based indices that included ACT scores as items (θ_2 and θ_5) had larger racial/ethnic differences than HSGPA (e.g., white students scored 0.80 SD higher than black students on θ_2, and 0.70 SD higher on HSGPA). The correlation of θ_5 and family income was 0.427, compared to 0.303 for the correlation of HSGPA and family income. The index that was based on the grouped school model with ACT scores for calibration (θ_6) also showed larger differences across socio-demographic groups. By grouping schools by mean achievement and calibrating with ACT scores, the GRM model produces ability estimates that vary considerably across school achievement groups, and by extension socio-demographic groups.

Table 4. Subgroup Differences

Measure	Black-white *d*	Hispanic-white *d*	Correlation with family income	Correlation with school mean ACT
HSGPA	-0.70	-0.53	0.303	0.285
HSAR index	-0.67	-0.48	0.300	0.279
GRM HSGPA-1 (θ_1)	-0.71	-0.51	0.313	0.279
GRM HSGPA-2 (θ_2)	-0.80	-0.55	0.372	0.374
GRM HSGPA-3 (θ_3)	-0.71	-0.51	0.315	0.281
GRM HSGPA-4 (θ_4)	-0.53	-0.42	0.241	0.130
GRM HSGPA-5 (θ_5)	-0.96	-0.62	0.427	0.524
GRM HSGPA-6 (θ_6)	-0.92	-0.61	0.392	0.476

4. Discussion

Educators, researchers, and policymakers alike have long stressed the importance of taking rigorous courses in high school to improve college readiness (Adelman, 1999; Adelman, 2006; Clinedinst & Koranteng, 2017; Gardner, Larsen, Baker, Campbell, & Crosby, 1983).

Based on the 2016 NACAC Admissions Trends survey, over half of colleges rated the strength of high school curriculum as a considerably important factor in college admission decisions (Clinedinst & Koranteng, 2017). The only factors rated as more important were grades in college prep courses, grades in all courses, and admission test scores. Despite the general consensus that rigor is important for promoting college readiness and success, how best to operationally define rigor remains an open question. This study contributes to the literature by developing GRM-based indices of high school coursework based on a large, representative sample of the general high school population and evaluating how well the indices predict degree completion as compared to ACT scores, HSGPA, and a previously derived prediction-based rigor index.

The findings of the current study highlight the benefit of using scaling-based methods to derive a weighted HSGPA as compared to indices based on optimized prediction. In particular, one strength of scaling-based methods is that they are not based on relationships to a specific outcome, or to any outcome for that matter. Therefore, if the desire is to create a rigor index that is predictive of multiple outcomes such as both first-year college GPA as well as degree completion, then scaling-based methods may be preferred. Indices based on optimized prediction may have the strongest relationship with the outcome on which it was derived but may exhibit weaker relationships with other outcomes. The results of the current study illustrate this point where the HSAR index, which was derived based on its relationship with first-year college GPA, was not as strongly related to degree completion as compared to five out of the six GRM-based indices. In fact, the correlation between bachelor's degree attainment and the HSAR index (r = .464) was quite a bit lower than that for θ_5 (.530), which exhibited the strongest relationship with earning a bachelor's degree.

The results of the current study also support the use of multiple measure models of college readiness. The GRM-based rigor indices added incrementally to the prediction of degree completion beyond traditional admissions measures, indicating that rigor provides unique information about a student's likelihood of future success. College and universities that consider multiple factors in college admission decisions, including the rigor of their high school coursework along with HSGPA and test scores, will have a more accurate picture of their applicants' level of college readiness and will be able to more precisely identify the students who are most likely to succeed. The results of the current study can help inform how college and universities derive a weighted HSGPA for their applicants to ensure the increase in prediction power.

We also found that the predictive strength of the rigor indices varied based on methodological decisions of: (1) whether ACT test scores were included in the model, and (2) whether high school courses were constrained to have the same difficulty and discrimination across schools. In general, the models that included ACT scores for both calibration and estimation had the largest correlations with degree completion (θ_2 and θ_5), followed by the models that included ACT scores only for calibration (θ_3 and θ_6), and lastly for the models that did not incorporate ACT scores in any way (θ_1 and θ_4). We also found that constraining high school courses to have the same difficulty and discrimination across all schools (θ_1, θ_2, and θ_3) resulted in lower validity coefficients than allowing the difficulty and discrimination to vary across groups of high schools (θ_4, θ_5, and θ_6). The indices that had the largest correlation with degree completion also exhibited the largest racial/ethnic subgroup differences. College and universities are often confronted with the competing goals of admitting the most qualified students while at the same time building a diverse class (Sackett, 2005). Therefore, colleges or universities interested in including rigor as an admission criterion may want to consider the index that best supports their mission and enrollment goals.

This study has many limitations worth noting. First, high school coursework and grade data were self-reported by the student during registration for the ACT. Even though research has shown that students tend to reliably report their coursework and grades when registering for the ACT (Sanchez & Buddin, 2015), it would have been preferable to use official high school transcript information to develop the rigor indices. However, those data were not available. With official high school transcript data, we would expect the predictive strength of the measures to improve somewhat (Kuncel, Credé, & Thomas, 2005). In a similar vein, limited response options are provided for the course grade information collected by ACT: A, B, C, D, and F. Students cannot report their grades at a finer level of granularity (e.g., A-, B+). Future research should examine whether the performance of the rigor indices can be improved when based on official transcript data which overcomes these challenges.

High schools were grouped by average ACT Composite score, and GRM-based indices treated courses from different groups as distinct. Grouping high schools resulted in increased predictive power; however, future research should examine how alternative grouping criteria, such as more fine-grain levels of ACT performance or by district or school, would affect the performance of the GRM-based indices. Reports of school-specific GRM parameters could be a resource for improving consistency across schools in grading standards and course difficulty.

Finally, the model specified a single ability estimate for each student based on four years of coursework and grade data as well as test scores. That a student’s ability is fixed across four years of high school may be an untenable assumption (see Hansen et al., 2016 for more discussion). Future research could examine models that treat ability as time-varying. Modeling course grades and test scores as functions of time-varying ability could yield a measure of academic momentum, which might have additional utility as a measure of college readiness.

In sum, the current study corroborates previous research highlighting the importance of rigor for college success. Scaling-based methods for producing summary measures of high school coursework and grades remain an attractive option for operationalizing rigor.

References

ACT (2014). *The ACT technical manual*. Iowa City, IA: ACT.

Adelman, C. (1999). *Answers in the tool box: Academic intensity, attendance patterns, and bachelor's degree attainment.* Washington, DC: U.S. Department of Education, National Institution on Postsecondary Education, Libraries, and Lifelong Learning.

Adelman, C. (2006). *The toolbox revisited: Paths to degree completion from high school through college.* Washington, DC: U.S. Department of Education.

Allen, J., Ndum, E., & Mattern, K. (2017). *An empirically-derived index of high school academic rigor*. Iowa City, IA: ACT.

Allen, J., Ndum, E., & Mattern, K. (2018). *An empirically-derived index of high school academic rigor*. Manuscript submitted for publication.

Allen, J., & Le, H. (2008). An additional measure of overall effect size for logistic regression models. *Journal of Educational and Behavioral Statistics, 33*(4), 416–441.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561-573.

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

Beatty, A. S., Sackett, P. R., Kuncel, N. R., Kiger, T. B., Rigdon, J. L., Shen, W., & Walmsley, P. T. (2012). *A comparison of alternate approaches to creating indices of academic rigor*. (College Board Research Report 2012-11). New York, NY: The College Board.

Clinedinst, M., & Koranteng, A. M. (2017). *2017 state of college admission.* Arlington, VA: National Association for College Admission Counseling. Retrieved from https://www.nacacnet.org/news--publications/publications/state-of-college-admission/.

Gardner, D. P., Larsen, Y. W., Baker, W., Campbell, A., & Crosby, E. A. (1983). *A nation at risk: The imperative for educational reform*. Washington, DC: U.S. Department of Education, the National Commission on Excellence in Education.

Hansen, J. D., Sadler, P. M., & Sonnert, G. (2016). *Watching our weights: Using a graded response model to estimate high school course rigor*. Retrieved from https://scholar.harvard.edu/files/john_hansen/files/grm_gpa.pdf.

Klopfenstein, K., & Lively, K. (2016). Do grade weights promote more advanced course taking? *Education Finance and Policy, 11*(3), 310-324.

Kuncel, N. R., Credé, M., & Thomas, L. L. (2005). The validity of self-reported grade point averages, class ranks, and test scores: A meta-analysis and review of the literature. *Review of Educational Research, 75*(1), 63-82.

National Student Clearinghouse. (2018). *Clearinghouse facts*. Retrieved from http://www.studentclearinghouse.org/about/clearinghouse_facts.php.

PrepScholar. (2018). *SAT/ACTPrep online guides and tips: What is a GPA scale? The 4.0 scale*. Retrieved from https://blog.prepscholar.com/what-is-a-gpa-scale.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *The Journal of the American Statistical Association, 82*, 387–394.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and Item Response Theory Analyses. *Journal of Statistical Software, 17*(5), 1-25. Retrieved from http://www.jstatsoft.org/v17/i05/.

Sackett, P. R. (2005). The performance-diversity tradeoff in admission testing. In W. J. Camara & E. W. Kimmel (Eds.), *Choosing students: Higher education admissions tools for the 21st century* (pp. 109-125). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Sadler, P. M., & Tai, R. H. (2007). Accounting for advanced high school coursework in college admissions decisions. *College and University Journal, 82*(4), 7-14.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34*, 100–114.

Samejima, F. (1997). Graded response model. In W. J. van der Linden, & R. K. Hambleton (Eds.) *Handbook of modern item response theory* (pp. 85-100). New York, NY: Springer.

Sanchez, E. I., & Buddin, R. (2015). *How accurate are self-reported high school courses, course grades, and grade point average?* Iowa City, IA: ACT.

SAS Institute Inc. (2008). *SAS/STAT® 9.2 user's guide*. Cary, NC: SAS Institute Inc.

Sawyer, R. (2010). *Usefulness of high school average and ACT scores in making college admission decisions*. Iowa City, IA: ACT.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3).

Wyatt, J. N., Wiley, A., Camara, W. J., & Proestler, N. (2011). *The development of an index of academic rigor for college readiness* (College Board Research Report 2011-11). New York, NY: The College Board.

Appendix

Table A1. Correlations and Summary Statistics

Variable	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.
1. HSGPA	1.000											
2. ACT Composite	0.614	1.000										
3. HSAR index	0.936	0.618	1.000									
4. GRM HSGPA-1 (θ_1)	0.969	0.657	0.928	1.000								
5. GRM HSGPA-2 (θ_2)	0.940	0.793	0.909	0.976	1.000							
6. GRM HSGPA-3 (θ_3)	0.966	0.660	0.927	1.000	0.977	1.000						
7. GRM HSGPA-4 (θ_4)	0.950	0.579	0.907	0.979	0.937	0.978	1.000					
8. GRM HSGPA-5 (θ_5)	0.911	0.828	0.880	0.941	0.981	0.942	0.870	1.000				
9. GRM HSGPA-6 (θ_6)	0.942	0.724	0.904	0.970	0.971	0.971	0.908	0.983	1.000			
10. Any degree/certificate	0.425	0.409	0.423	0.432	0.455	0.432	0.389	0.475	0.465	1.000		
11. Bachelor’s or higher	0.459	0.468	0.464	0.477	0.509	0.478	0.428	0.530	0.515	0.841	1.000	
12. Post-bachelor’s	0.183	0.194	0.193	0.205	0.218	0.207	0.187	0.220	0.214	0.225	0.267	1.000
Mean	3.082	20.851	1.754	0.035	-0.006	-0.005	0.115	-0.015	0.000	0.468	0.384	0.043
Standard Deviation	0.638	5.186	0.538	0.928	0.910	0.924	0.879	0.955	0.977	0.498	0.485	0.202
Skewness	-0.570	0.341	-0.475	0.074	0.220	0.140	0.073	0.135	0.069	0.126	0.476	4.529

Table A2. Predictors of College Degree Attainment (Bachelor's Degree or Higher)

Predictor	Model number / adjusted odds ratio (95% confidence interval)							
	1	2	3	4	5	6	7	8
HSGPA	2.87 (2.77,2.97)	1.77 (1.48,2.12)	1.76 (1.57,1.97)	1.81 (1.63,2.01)	1.83 (1.64,2.04)	2.48 (2.24,2.74)	1.27 (1.15,1.39)	1.27 (1.16,1.40)
ACT Composite	1.63 (1.58,1.69)	1.57 (1.51,1.63)	1.55 (1.49,1.60)	1.33 (1.26,1.41)	1.55 (1.50,1.60)	1.61 (1.55,1.66)	1.12 (1.06,1.18)	1.48 (1.42,1.53)
HSAR index		1.70 (1.40,2.06)						
GRM HSGPA-1 (θ_1)			1.61 (1.45,1.79)					
GRM HSGPA-2 (θ_2)				1.82 (1.59,2.07)				
GRM HSGPA-3 (θ_3)					1.56 (1.41,1.72)			
GRM HSGPA-4 (θ_4)						1.15 (1.06,1.26)		
GRM HSGPA-5 (θ_5)							3.14 (2.76,3.57)	
GRM HSGPA-6 (θ_6)								2.36 (2.14,2.61)
Model *OOR* (bachelor's +)	3.92	4.04	4.05	4.07	4.05	3.96	4.15	4.12
Model *OOR* (any degree)	3.03	3.11	3.13	3.14	3.14	3.05	3.20	3.20
Model *OOR* (post-bachelor's)	3.17	3.17	3.13	3.14	3.14	3.20	3.10	3.11

Note: *OOR* = overall odds ratio for multiple logistic regression model