RESEARCH REPORT 2012-11

CollegeBoard

A Comparison of Alternate Approaches to Creating Indices of Academic Rigor

By Adam S. Beatty, Paul R. Sackett, Nathan R. Kuncel, Thomas B. Kiger, Jana L. Rigdon, Winny Shen, and Philip T. Walmsley

Adam S. Beatty conducted this research as a doctoral student in Industrial and Organizational Psychology at the University of Minnesota. He is now a research scientist at the Human Resources Research Organization (HumRRO).

Paul R. Sackett is the Beverly and Richard Fink Distinguished Professor of Psychology at the University of Minnesota.

Nathan R. Kuncel is the Marvin D. Dunnette Distinguished Professor of Psychology at the University of Minnesota.

Thomas B. Kiger conducted this research as a doctoral student in Industrial and Organizational Psychology at the University of Minnesota. He is now a research scientist at the Human Resource Research Organization (HumRRO).

Jana L. Rigdon is a doctoral student in Industrial and Organizational Psychology at the University of Minnesota.

Winny Shen conducted this research as a doctoral student in Industrial and Organizational Psychology at the University of Minnesota. She is now an assistant professor of psychology at the University of South Florida.

Philip T. Walmsley is a doctoral student in Industrial and Organizational Psychology at the University of Minnesota.

About the College Board

The College Board is a mission-driven not-for-profit organization that connects students to college success and opportunity. Founded in 1900, the College Board was created to expand access to higher education. Today, the membership association is made up of over 6,000 of the world's leading educational institutions and is dedicated to promoting excellence and equity in education. Each year, the College Board helps more than seven million students prepare for a successful transition to college through programs and services in college readiness and college success — including the SAT® and the Advanced Placement Program®. The organization also serves the education community through research and advocacy on behalf of students, educators and schools. For further information, visit www.collegeboard.org.

RESEARCH

For more information on College Board research and data, visit research.collegeboard.org.

Contents

Tables

Executive Summary

In recent decades, there has been an increasing emphasis placed on college graduation rates and reducing attrition due to the social and economic benefits, at both the individual and national levels, proposed to accrue from a more highly educated population (Bureau of Labor Statistics, 2011). In the United States in particular, there is a concern that declining college graduation rates relative to the rest of the world's population will reduce economic competitiveness (Callan, 2008). As such, in addition to research on how to increase educational performance in elementary and secondary schools, educational researchers are also interested in the determinants of performance and persistence at the collegiate level.

One method hypothesized to promote increased college graduation rates is to raise the standards in the nation's high schools in order to better prepare students for college. Indeed, data from many converging sources suggests high school graduates are not prepared for higher-level college curriculum (Achieve, 2005). As such, many state institutions have attempted to set standards for rigor in order to ensure students are prepared for college study. Against this backdrop, the College Board has recently developed a measure of academic rigor, termed the Academic Rigor Index (ARI), for the purpose of examining how well a student is prepared for college study both within and across broad content domains (Wiley, Wyatt, & Camara, 2010). The ARI awards 0 to 5 points in each of five areas (English, mathematics, science, social science/history, and foreign/classical languages) based on students' self-reported course-taking and sums these to create an overall index on a 0–25-point scale. The 25 credited activities are drawn from a larger set of course-taking variables. Each individual credited activity's inclusion in the index is empirically supported by links to subsequent collegiate performance. The decisions to award an equal number of possible points in each of the five areas, and to weight each area equally in computing the total score, were not empirically based, and thus the degree to which relaxing the equal point per area and equal weight per area constraints could improve the predictive power of the ARI is not known. The purpose of the present paper is to compare the ARI with alternative scoring procedures that remove these constraints.

Introduction

Rational vs. Empirical Scoring Procedures

In the psychological test development process, two main classes of methods by which to select items and develop scoring keys can be considered: rational and empirical. A test developer who subscribes to a purely rational point of view would suggest that items should (1) be selected on the basis of their judged relevance to the outcome of interest (e.g., job or academic performance, the absence of illegal behaviors at work, etc.), and (2) that their response options should be assigned a value based on the hypothesized relationship with that outcome. Both of these decisions will often be driven by theory. A "rational" test-developer may suggest that, for instance, when developing a test for a software engineer, it is standard that an applicant would have taken at least five computer science courses while completing a bachelor's degree, so this would be worth 0 points, whereas each course taken beyond that would be worth 1 additional point, and this would be the scoring key applied to that item. This entire process can be conducted using the developer's expert judgment and understanding of the content domain.

At the other extreme, a test-developer subscribing to a purely empirical point of view would be likely to gather a large pool of items thought to be at all relevant to the outcome of interest. They would then administer this pool of items to a sample of members of the class of interest (e.g., current employees, current graduate students), along with gathering outcome data (e.g., grades, retention). Items would then be selected based on the strength of their relationship to that outcome, and the value given to each item response would be derived through a statistical procedure, also based on its relationship to the outcome. As a result, it is an almost entirely data-driven process, often requiring the use of a large number of research subjects and cross-validation procedures.

A key aspect of the debate between these two points of view is a trade-off between validity and construct purity/meaningfulness. Specifically, weights derived through an empirical least-squares minimization process represent optimal weights for the sample and given a representative and large enough sample, the degree to which these weights would exhibit shrinkage should be minimal. While many implementations of empirical-keying do not use procedures that produce optimal weights, validity has generally been found to be solid (Hogan, 1994). Proponents of the rational method tend to argue that the goal should be to identify constructs and determine the validity of these constructs, rather than individual items. As empirical methods can potentially result in items being weighted in the opposite direction from what would be rationally expected, this can make finding an overarching narrative for *why* the empirical key works difficult.

In practice, most methods of construct-related test development represent a mixture of the rational and empirical paradigms. Items derived rationally may be removed if there is no empirical relationship with the outcome when used operationally, and item responses weighted heavily empirically may appear unintuitive and dropped for practical reasons. While this mixture of methods has been acknowledged in some sources, the distinction between rational and empirical paradigms represents somewhat of a schism within researchers in applied psychology, particularly with respect to personality and biographical data (biodata) inventories.

Academic Rigor Index

The ARI was developed by analyzing a large (*N* = 67,644) data set that contained detailed high school course-taking information, SAT®, high school GPA, and college first-year GPA (FGPA) and retention criteria (Wiley et al., 2010). Like many construct-oriented inventories, its scoring rules are predicated on a mixture of rational and empirical concerns. First, the ARI contains components of the College Board's definition of a core high school curriculum (four years of English, three years of math, three years of natural science, and three years of social science/history). Additionally, a number of comparisons were conducted such that the mean FGPA of those who participated in a particular high school course, number of courses in a domain, or Advanced Placement® (AP®)/honors option of a course was compared to those who did not, and courses that resulted in a mean FGPA difference of students greater than .05 were considered to be meaningful for rigor. These could be said to be loosely keyed on FGPA. Finally, these FGPA-comparison identified rigor variables were combined and reduced in such a way as to create five domain-level scales, each with 5 points. As a result, this process can be described as a combination of an empirical and a rational approach: empirical in that each score point reflects a course-taking activity that is empirically linked to FGPA, and rational in the decision to award equal points per domain and to weight the five domains equally. For more detail on how the ARI was developed, see Wiley et al. (2010).

The ARI development procedure provides at least two areas where less than optimal results can occur. First, the rational decision rules used could utilize only a portion of the variance in the academic rigor items related to the criteria. As these decision rules were grounded in the empirical data, large departures from optimality seem unlikely. However, it seems useful to examine the extent to which, if any, empirical-keying methods produce more valid weights. Second, given the nature of the data-collection procedure (i.e., being tied to the administration of the SAT), incomplete response patterns could result based on when students completed the academic rigor items. Specifically, if students completed the SAT at some point in 11th grade, they may not know which courses they would be taking in 12th grade and, therefore, left these items blank. For ARI points that are determined by number of years of course-taking within a certain discipline, this results in the potential to make earning these points unlikely or impossible. An empirical-keying method at the item level would allow these students to receive "partial credit" based on the number of courses they actually took.

There are also reasons to expect that the specific weighting method used will not make a large difference. Perhaps most prominent of these is a research stream that suggests both that the precision of the weights used does not create wildly different results and that as long as the directional sign is accurate, weights can be unit or rounded and exhibit similar results to the least-squares weights (e.g., Green, 1977; Rozeboom, 1979; Wilks, 1938). In smaller samples, these alternative weights can even outperform least-squares weights in cross-validation (e.g., Dawes & Corrigan, 1974; Einhorn & Hogarth, 1975; Raju, Bilgic, Edwards, & Fleer, 1997; Schmidt, 1971). This suggests that differences between a process that relies on data to some degree to establish its weights and a process that relies entirely on empirical data could be negligible.

Given the importance of the broader goal of college success, it is critical that the predictive power of academic rigor be as close to purely empirical methods, or as high as possible. Overall then, the goal of our study is to compare the rational/empirical method of scoring the academic rigor items used in developing the ARI to strictly empirical methods and different levels of aggregation of the academic rigor items.

Method

Sample

The sample used in this study is a subset of the SAT Admitted Class Evaluation Service™ (ACES™) data set developed by the College Board in collaboration with a large group of universities and colleges. The colleges were purposefully chosen to be broadly representative of regions, large and small schools, and public and private schools. The sample used includes the 67,644 students in the ACES data set for the entering class of 2007 for 110 schools who fully completed questions on all relevant measures of interest. Descriptive comparisons between this subset and the full data set of students in the ACES data set suggested that distributions of variables were fairly similar in both data sets, though more sophisticated missing data techniques were not conducted (Wiley et al., 2010).

Measures

Freshman Grade Point Average. Freshman grade point average (FGPA) was provided by the colleges/universities after the student's first year at that institution.

Retention. Retention was also provided by the college or university and was a dichotomous variable indicating whether the student had returned to that institution for a second year.

High School Grade Point Average. The measure of high school GPA (HSGPA) used in this study was a self-report from a questionnaire students filled out at the time that they took the SAT®. While there are bound to be differences between self-report measures and institution reported measures (e.g., due to rounding, forgetfulness, school-weighting or recomputation or willful distortion), Kuncel, Credé, & Thomas (2005) found a mean $r = .82$ between self-reported and high-school-reported GPA. In addition, previous work with College Board data sets has reported that the correlation between these two variables is .75 (Beatty et al., 2010). As such, it seems reasonable to expect a high degree of overlap and similar correlational results.

SAT. SAT scores were collected by the College Board. Scores on the three SAT subtests (Math, Critical Reading, and Writing) were combined into a single unit-weighted composite for use within our study.

Academic Rigor. Questions related to high school curriculum were asked with respect to five academic domains at the time of taking the SAT: English, mathematics, natural science, social science/history, and foreign and classical languages. Course work information obtained included general course titles and grade level when taken as well as AP, honors, or dual-enrollment participation. Each question required a simple Yes/No based on whether a student participated in a given course in a given grade and there were 395 items in total.

Procedures and Data Analysis Plan

Our first task was to recreate the College Board's Academic Rigor Index (Wiley et al., 2010). This process involves a straightforward simplification and coding of the 395 academic rigor items. Specifically, for each of five domains (English, Math, Science, Social Science, and Foreign Language), 5 points were allocated based on aggregated student-level data, such as the number of courses taken in that domain and whether any (or how many) AP, honors, or dual-enrollment courses were taken. This resulted in a 25-point scale. As our data set was the same data set used by Wiley et al. (2010) to explore the properties of the ARI, we were able to directly compare the results of our implementation of their scoring rules in order to ensure

we had done so correctly. For more detail on the development, scoring, and properties of the ARI, see Wiley et al. (2010).

Empirical Scoring Procedures

The scoring procedure for the ARI contains a number of tacit assumptions. First and foremost, there is the assumption that a much simpler, aggregated representation of the 395 academic rigor items is adequate to capture their variance (e.g., swapping the number of foreign language courses taken for the dozens of specific language/grade combinations doesn't lose much in the way of prediction). Second, the scoring rule of allocating 5 points per domain suggests an equivalent contribution of academic rigor in each domain to criterion performance. Likewise, allocating one point to each behavior of interest (rather than differential regression weights, for instance) suggests that each of these behaviors is equally relevant for criterion performance. These assumptions can be considered to be nested hierarchically (i.e., items within aggregated points within domain), with each level adding a new constraint.

While there are good reasons for expecting that many of these assumptions would hold, it seems reasonable to investigate how much precision in prediction is lost by each of these assumptions. As such, we applied two methods of empirical keying to models that represent each of these assumptions.

Vertical Percent Method

The vertical percent method is a classical empirical-keying procedure often used in the domain of biographical data (biodata). The basic procedure and weighting scheme was introduced by Strong (1926), and later outlined by England (1961, 1971). Hogan (1994) suggests that it is the most common method used to empirically key biodata inventory items, and it has a considerable amount of evidence supporting its use (e.g., Devlin, Abrahams, & Edwards, 1992; Mitchell & Klimoski, 1982).

While implementing the vertical percent method can be a computationally intensive process, especially when compared to other procedures in the modern era, it is also straightforward. First, high and low criterion groups are chosen. When using dichotomous criteria (e.g., retention), this is simple, with those who returned to college for the second year being coded as the high criterion group, and those who did not return coded as the low criterion group. When using continuous criteria, such as GPA, the cut score to determine high and low group membership is up to the discretion of the researcher. Hogan (1994) suggested that a common procedure was fashioned after Kelley (1939) and was to set the high criterion group as those who score in the upper 27% on the criterion, and the low criterion group as those who score in the lower 27%. Devlin et al. (1992) found that results were fairly stable across a number of different splits (e.g., top and bottom 5%, top and bottom 10%, etc.) in predicting a GPA analogue for the U.S. Naval Academy. Additionally, in previous research with a closely related College Board data set, we tested multiple splits (1%, 5%, 10%, 15%, 20%, and 27%) and found that the resulting composites all correlated very highly and thus did not affect conclusions (Beatty et al., 2011). In this study, we chose the top and bottom 20% on the FGPA criterion (above 3.57 and below 2.39, respectively).

After criterion groups are chosen, the next step involves computing the percent within each of the criterion groups that chose each item alternative. Then, the percent from the low criterion group is subtracted from the percent in the high criterion group. This percentage difference is applied to each of the subjects' item responses as a weight, and these weights are summed to create a total composite for each subject.

Since empirical-keying processes rely exclusively on the data to derive weights, these weights are optimized for that specific sample, and the validity of any such composite would be expected to exhibit shrinkage when applied to a new sample. As a result, even though we had a large and arguably representative sample of students, we used a cross-validation procedure to test the applicability of the weights derived from the vertical percent method. Specifically, we randomly selected two-thirds of the entire sample of students to be in the weight-development group, and then assigned the remaining one-third of the sample to be in the cross-validation group. This resulted in sample sizes of 44,701 and 22,943, respectively, in the analyses involving FGPA, and 40,413 and 20,767 across 101 schools in the analyses involving retention. Both the number of schools and total *N* differ in the retention analyses due to some schools having either no, or very small, reported attrition, resulting in a lack of variability.

Weights derived in the weight-derivation group were applied to both the weight-derivation and cross-validation groups, and composites were computed. These composites were then used to predict FGPA and retention in order to obtain the validity of each individual composite. Incremental validity over HSGPA and SAT was then tested by hierarchical regression. The measure of variance explained reported for logistic regression in this study is the Nagelkerke pseudo-R^2 (1991). Finally, descriptive statistics and regressions were calculated both within each school and *n*-weighted and across the total sample in order to come up with two sets of estimates. We choose to focus on results computed within-school and aggregated as this reflects solely on within-school effects and does not conflate estimates with school-level effects. However, the total sample results are presented in an appendix for comparison purposes with the Wiley et al. (2010) report.

Multiple Regression

Another traditional method for deriving weights for a linear composite is regression. While the vertical percent method examines each item and its response options in isolation from any other item, regression examines the interrelations between the items and assigns weights based on the entire item set. Results from regression and the vertical percent method would thus be expected to differ based on the amount of intercorrelation within the data such that regression could pick up redundancy between the items. Beatty et al. (2010) recently compared the two procedures and found that regression produced more cross-valid weights when the sample size approached approximately 1.75 times the number of items, and that as the sample size increased, it had the potential to produce weights that had substantially more validity than the vertical percent method (in some cases, doubling the validity). As the ARI items are all proposed to reflect a common construct, it seems likely that item intercorrelations would be high, and therefore, allow potential for multiple regression to provide superior weights relative to the vertical percent method. As the items are all dichotomous, they are already equivalent to dummy-coding, and can be directly used in regression. We therefore regressed FGPA and retention on the composites, and also examined the incremental validity of the subsequent multiple regression-derived composites over HSGPA and SAT, both on the total sample and n-weighting results across schools. The same cross-validation procedure was used as described previously, and the samples used in each group were based on the same random draw.

Models Tested

Both the vertical percent method and multiple regression were used to evaluate three models corresponding to the assumptions discussed above. First, in order to examine the assumption that the simplified set of variables in the ARI accounts for a large portion of the variance in

the 395 academic rigor items, we simply applied the vertical percent method and multiple regression to the entire set of items. Second, to test whether any predictive efficacy was given up by weighting each of the ARI's 25 points equally, we applied the vertical percent method and multiple regression to these 25 points (i.e., the effect of unit weights versus differential weights). Third, we examined the importance of weighting academic content domains differently by using these two procedures on the five domain composite variables.

Additionally, it is important to note that regression composites will always explain more variance in the criterion for every additional item entered into the composite. In order to determine whether any advantage from using the full 395-item regression composite is solely a function of there being more items than in the ARI, we also computed a composite of the most predictive 25 items identified through forward stepwise regression. Though many previous methodologists have suggested that stepwise regression is not an ideal procedure (e.g., Copas, 1983; Leigh, 1988; Thompson, 1995), we chose to utilize it because other readily available metrics for inclusion (e.g., standardized and unstandardized regression weights) identified variables that had heavy weight when they were endorsed but were quite rare, resulting in composites with very little variability.

Results

We preview the results to be presented in detail below. An initial finding is that the vertical percent method was never superior to the regression method. Thus the substantive discussion that follows focuses on the regression results. Table 1 addresses the central question for the current study showing the difference in the predictive power of the ARI and empirical regression-based models for both FGPA and retention. Further, incremental validity over and above HSGPA and SAT is displayed. The ARI composite (r = 0.24) performs nearly as well as the regression model utilizing all 395 items (r = 0.26) in the prediction of FGPA in the cross-validation sample. When looking at retention the ARI composite again predicts almost as well as the regression-based models. In addition, little difference in terms of incremental validity was found. We note that we use the traditional validity coefficient for presenting FGPA results and present results in the variance explained metric for retention as we argue it is more appropriate to focus on results derived from a logistic framework for dichotomous criteria.

Tables 2 and 3 provide more detailed information displaying means, standard deviations, and correlations between study variables for both the weight-derivation and cross-validation samples. These tables report results relevant to the prediction of FGPA and retention, respectively, calculated within-school and sample-size weighted. Given the large sample sizes involved, it is perhaps not surprising that there doesn't appear to be a large amount of shrinkage involved for most composites (the most prominent exception being the full 395-item retention logistic regression). Also not surprising are the lower correlations for the prediction of retention relative to FGPA. The content domain composites keyed to each criterion tend to be highly correlated with each other (e.g., .60s–.90s).

Tables 4 and 5 present the results of the regression analyses of each of the composites. Table 4 reports results for the prediction of FGPA sample size-weighted across school. The full set of 395 items analyzed by regression provides the most variance explained out of the composites predicting FGPA alone; however, the absolute differences between the composites appear quite small. To aid in interpretation, percentages of the variance explained of the full-set regression composite are also reported. While regression involving the full item set holds a sizeable edge in the weight-derivation sample, this is substantially reduced when these weights are applied to the cross-validation sample (e.g., many composites exhibit up to and over 90% of the maximum variance explained). A similar pattern of results between the composites holds for the investigations of incremental variance, however, the absolute value of all of these composites in incrementing over HSGPA and SAT for the prediction of FGPA is small (e.g., change in $R^2 < .01$).

Interestingly, the regression results support the earlier hypothesis that the empirical models represent a series of constraints: from the rational/empirical ARI key, to the domain-factor empirical key, to the 25-point empirical key, to the full-item empirical key, each result in progressively more variance explained. However, this is not the case for the vertical percent method. While the 25-point key is more predictive than the domain-factor key, the jump to the 395-item key results in a decrement to its ability to predict FGPA. This is consistent with Beatty et al.'s (2010) theory for why weights derived via multiple regression can exceed those derived by the vertical percent method. Specifically, the inclusion of hundreds of related pieces of information increased the redundancy within the items: redundancy that the vertical percent method does not account for in its weighting.

Table 5 provides results for these same scoring models in predicting retention for sample-size weighted school-level results. Results mirror those for FGPA, with two major exceptions.

First, all models performed very similarly relative to the full item-set regression composite, and there was not an instance where certain tests lagged behind the others as there was with FGPA. Second, the composites exhibit much more of an increment to HSGPA and SAT in predicting retention, both in absolute terms and in percentage of the variance explained by HSGPA and SAT. This suggests that even though the ARI and its components were keyed on FGPA, they may have more utility for the prediction of retention.

The appendix contains analogues for each of the above discussed tables, but with analyses conducted on the entire sample to allow comparison to results presented by Wiley et al. (2010). The pattern of findings is almost identical in both sets of analyses.

Discussion

The goal of this study was to examine different methods of scoring raw data pertaining to academic rigor in order to ascertain the resulting effect on both cross-valid variance explained, and incremental validity over HSGPA and SAT in predicting college FGPA and retention. Overall, results suggest that the rational/empirical method used in developing the ARI resulted in little information loss. While there were larger differences between the methods in the weight-derivation sample, these differences tended to shrink dramatically in the cross-validation sample. Although the incremental validity of any academic rigor composite over HSGPA and SAT tended to be small in an absolute sense (e.g., approximately 1% of variance explained), it appeared to be relatively more predictive for retention, which is a notoriously difficult criterion to predict.

First, these results have relevance for the utility of academic rigor in predicting educational criteria. Academic rigor proved to be a fairly strong predictor of first-year college grades. Although the incremental validity above traditional predictors was not large, even small gains can be beneficial. Further, it provides college admission offices with more unique and detailed information about students. This could aid decisions for more specific domains and serve as an indicator of interest and/or motivation. More notably, academic rigor shows great promise in addressing current societal goals to aid college retention and completion rates. The use of an academic rigor composite results in relatively large gains of incremental validity over traditional predictors. Given that the information that goes into the academic rigor composite is routinely provided to colleges, this seems to be a plausible step for increasing retention rates, either by selecting those who have a higher likelihood of retention or identifying those who are of greater risk to attrite.

Second, this study's findings provide a straightforward example of well-known psychometric results in the weighting of linear composites. Research reviewed earlier indicated that validity results were often insensitive to the weighting strategy used. In accord with this, there were very few differences among the weights. In addition, a subset of the analyses was representative of an increasing order of generality (from the rational/empirical keying method used in the ARI, to the domain-factor empirical key, to the 25-point empirical key, to the full-item empirical key). This pattern of results was expected given the sample size, though it is perhaps surprising how little the levels of generality and differential weighting mattered. Of course, it is plausible that regression and vertical percent weights at different levels of generality would exhibit worse validity than the rational/empirical ARI composite with small sample sizes, but for a fair comparison of these procedures, the rational/empirical ARI composite would also need to be re-created based on the FGPA comparisons utilized to help generate its weights.

While any weighting scheme can be computationally automated, test use occurs in a social context. As a result, an additional consideration in favor of the rational/empirical ARI scoring method is its explanatory simplicity and face-validity. The 395-item regression composites contain a number of examples where taking a certain course in 9th grade is weighted positively, whereas taking it in 12th grade is weighted negatively, or where the highest weights are in relatively rare course-options at the high-school level (e.g., Greek, Russian, and Korean). Having to defend these decision rules in a high-profile inventory could be associated with public relations issues and negative test-taker reactions, and doing so only seems justified if they provide substantially more explanatory power by capturing critical but unintuitive relationships. As that does not appear to the case, the ARI-scoring key appears to balance these test-use and validity concerns.

Finally, these results have relevance for the more general domain of empirical-keying. Previous research has indicated that multiple regression's ability to determine valid item weights exceeds that of the vertical percent method (Beatty et al., 2010). The proposed mechanism by which this occurred was through regression's ability to account for redundancy among variables. The results of this study suggest a similar inference, with the vertical percent method resulting in similar validities to regression until the full 395-item data set where one would expect the most redundancy. As such, this study adds a converging line of evidence to this research stream. As for the broader issue of a comparison of the validity between rational and empirical methods, this study isn't particularly illuminating. While we note that the ARI has some rational elements in its construction, it appears to be heavily based on an empirical examination of the relationship of high school course data to FGPA, and Wiley et al. (2010) themselves refer to it as an empirical key.

In the technical report describing the development of the ARI, Wiley et al. (2010) suggest a few limitations and areas for additional research. These include the need to compare the self-reported high school course work data and actual transcript data to investigate their convergence, and the necessity of relying on broad high school course titles along with an assumed equivalence of difficulty within them (e.g., honors courses at different institutions could indicate more or less rigor). We concur that these would be useful investigations, and that they could inform any subsequent research conducted with the ARI. In addition, we suggest that the ARI, or its more specific components, may have greater efficacy and incremental validity in predicting specific-domain course grades rather than broad FGPA and could thus serve as a proxy for engagement in the domain.

Overall, a comparison of a number of different methods and approaches to computing composites of high school course data in order to estimate academic rigor suggested few differences in cross-valid variance explained. Results confirmed that the College Board's Academic Rigor Index gives up little validity in its coding of high school course data.

References

Achieve Inc. (2005). *Rising to the challenge: Are high school graduates prepared for college and work? A study of recent high school graduates, college instructors, and employers.* Retrieved from http://www.achieve.org/files/pollreport_0.pdf

Beatty, A. S., Sackett, P. R., Kuncel, N. R., Rigdon, J. L., Shen, W., & Kiger, T. B. (2010). *Testing the limits of empirically based prediction of college freshman grade point average and retention using information from the student descriptive questionnaire.* Research Report submitted to the College Board, New York, NY.

Beatty, A. S., Sackett, P. R., Kuncel, N. R., Rigdon, J., Shen, W., & Kiger, T. B. (2011, April). *A comparison of two methods for keying biodata inventories.* Poster presented at the annual meeting of the Society for Industrial and Organizational Psychology, Chicago, IL.

Bureau of Labor Statistics. (2011). *Occupational outlook handbook, 2010–2011 Edition*. Retrieved from http://www.bls.gov/oco

Callan, P. M. (2008). The 2008 national report card: Modest improvements, persistent disparities, eroding global competitiveness. Retrieved from http://measuringup2008.highereducation.org/commentary/callan.php

Copas, J. B. (1983). Regression, prediction and shrinkage. *Journal of the Royal Statistical Society. Series B (Methodological), 45*(3), 311–354.

Dawes, R.M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81*, 95–106.

Devlin, S. E., Abrahams, N. M., & Edwards, J.E. (1992). Empirical keying of biographical data: Cross-validity as a function of scaling procedure and sample size. *Military Psychology, 4*, 119–136.

Einhorn, H. J., & Hogarth, R. M. (1975). Unit weighting schemes for decision making. *Organizational Behavior and Human Performance, 13*, 171–192.

England, G. W. (1961). *Development and use of weighted application blanks.* Dubuque, IA: Brown.

England, G. W. (1971). *Development and use of weighted application blanks* (Bulletin No. 55), Minneapolis: University of Minnesota, Industrial Relations Center.

Green, B. F., Jr. (1977). Parameter sensitivity in multivariate methods. *Multivariate Behavioral Research, 12*(3), 263–287.

Hogan, J. B. (1994). Empirical keying of background data measures. In G. S. Stokes, M. D. Mumford, & W.A. Owens (Eds.), *Biodata handbook: Theory, research, and use of biographical information in selection and performance prediction* (pp. 69–107). Palo Alto, CA: Consulting Psychologists Press.

Kelley, T. L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology, 30*, 17–24.

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.

Leigh, J. P. (1988). Assessing the importance of an independent variable in multiple regression: Is stepwise unwise? *Journal of Clinical Epidemiology, 41*(7), 669–677.

Mitchell, T. W., & Klimoski, R. J. (1982). Is it rational to be empirical? A test of methods for scoring biographical data. *Journal of Applied Psychology, 67*(4), 411–418.

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika, 78*, 691–692.

Raju, N. S., Bilgic, R., Edwards, J. E., & Fleer, P. F. (1997). Methodology review: Estimation of population validity and cross-validity, and the use of equal weights in prediction. *Applied Psychological Measurement, 21*(4), 291–305.

Rozeboom, W.W. (1979). Sensitivity of a linear composite of predictor items to differential item weighting. *Psychometrika, 44*(3), 289–296.

Schmidt, F.L. (1971). The relative efficiency of regression and simple unit predictor weights in applied differential psychology. *Educational and Psychological Measurement, 31*(3), 699–714.

Strong, E. K., Jr. (1926). An interest test for personnel managers. *Journal of Personnel Research, 5,* 194–203.

Thompson, B. (1995). Stepwise regression and stepwise discriminant-analysis need not apply here: A guidelines editorial. *Educational and Psychological Measurement, 55*(4), 525–534.

Wiley, A., Wyatt, J. & Camara, W. J. (2010). *The development of a multidimensional index of college readiness for SAT students.* (College Board Research Report 2010-3). New York, NY: The College Board.

Wilks, S. S. (1938). Weighting systems for linear functions of correlated variables when there is no dependent variable. *Psychometrika, 3*(1), 23–40.

Table 1.

Summary of Comparisons of Different Methods for Predicting FGPA and Retention – *N*-Weighted by School

	r with FGPA	R^2 with Retention	Incremental R^2 for FGPA	Incremental R^2 for Retention
ARI Composite	.256 (.243)	.019 (.021)	.005 (.005)	.009 (.014)
Regression – all 395 items	.299 (.264)	.039 (.023)	.010 (.007)	.025 (.016)
Regression – 25 points	.263 (.253)	.022 (.022)	.005 (.006)	.010 (.013)
Regression – 5 domain factors	.263 (.251)	.020 (.021)	.005 (.006)	.009 (.014)

Note. Incremental variance explained is over HSGPA and SAT. For all cells, the first value is for the weight-development sample, whereas the value in parentheses is for the cross-validation sample. Statistics presented in this table are *n*-weighted means of results from 67,644 students in 110 schools for the FGPA criterion, and 61,180 students in 101 schools for the retention criterion.

Table 2.

Descriptive Statistics and Correlations Between Variables Used in Predicting FGPA – *N*-Weighted by School

	Mean[a]	SD[a]	1	2	3	4	5	6	7	8	9	10	11
1. FGPA	2.93 (2.93)	.67 (.66)		.38	.37	.26	.26	.26	.30	.26	.23	.26	.25
2. HSGPA	3.62 (3.61)	.43 (.44)	.37		.27	.30	.30	.30	.28	.29	.25	.31	.31
3. SAT	1662.7 (1660.7)	204.6 (204)	.36	.26		.46	.46	.45	.48	.45	.40	.46	.44
4. ARI Composite	.62 (.62)	.21 (.21)	.24	.29	.45		.95	.92	.75	.74	.69	.98	.92
5. Regression – 5 Domain Factors	.65 (.65)	.22 (.22)	.25	.28	.46	.95		.96	.78	.76	.71	.94	.91
6. Regression – 25 points	.48 (.48)	.22 (.22)	.25	.28	.44	.92	.96		.78	.77	.70	.94	.91
7. Regression – Full 395 Items	.37 (.37)	.25 (.25)	.26	.27	.47	.75	.78	.78		.86	.73	.76	.75
8. Regression – Best 25 Items	.20 (.20)	.23 (.23)	.24	.28	.43	.74	.76	.77	.86		.76	.77	.76
9. Vertical Percent Method – Full 395 Items	-2.58 (-2.56)	3.17 (3.16)	.21	.23	.40	.69	.72	.70	.73	.75		.72	.71
10. Vertical Percent Method – 25 points	-.15 (-.16)	2.31 (2.30)	.25	.29	.45	.98	.94	.94	.76	.77	.73		.95
11. Vertical Percent Method – 5 Domain Factors	-.02 (-.02)	.47 (.46)	.24	.29	.43	.91	.90	.91	.74	.76	.71	.95	

Note. Correlations above the diagonal are correlations from the weight-development sample. Correlations below the diagonal are from the cross-validation sample. All correlations are significant at $p < .001$. All statistics presented in this table are *n*-weighted means of results from 67,644 students in 110 schools.

a. First value is for weight-development sample. Values in parentheses are for the cross-validation sample.

Table 3.

Descriptive Statistics and Correlations Between Variables Used in Predicting Retention – *N*-Weighted by School

	Mean[a]	SD[a]	1	2	3	4	5	6	7	8	9	10	11
1. HSGPA	3.62 (3.61)	.43 (.44)		.26	.08	.29	.28	.27	.21	.24	.23	.30	.30
2. SAT	1663.4 (1662.1)	201.3 (200.7)	.25		.06	.45	.45	.42	.35	.40	.38	.45	.43
3. Retention	.87 (.86)	.32 (.32)	.08	.06		.07	.08	.08	.12	.09	.07	.07	.07
4. ARI Composite	1.37 (1.37)	.46 (.46)	.28	.44	.06		.94	.89	.61	.69	.66	.98	.91
5. Regression – 5 Domain Factors	1.48 (1.48)	.47 (.47)	.26	.45	.06	.94		.94	.63	.72	.69	.93	.90
6. Regression – 25 points	.95 (.95)	.49 (.49)	.26	.42	.06	.89	.94		.64	.75	.67	.92	.90
7. Regression – Full 395 Items	.76 (.75)	.72 (.74)	.21	.36	.06	.62	.64	.65		.74	.57	.62	.61
8. Regression – Best 25 Items	.62 (.62)	.55 (.55)	.24	.40	.06	.70	.73	.75	.74		.68	.72	.72
9. Vertical Percent Method – Full 395 Items	-1.35 (-1.33)	1.72 (1.72)	.21	.38	.05	.67	.69	.67	.57	.68		.70	.69
10. Vertical Percent Method – 25 points	-.05 (-.05)	1.23 (1.23)	.29	.44	.06	.98	.93	.92	.63	.72	.71		.95
11. Vertical Percent Method – 5 Domain Factors	-.01 (-.01)	.25 (.25)	.28	.42	.06	.91	.90	.90	.62	.72	.69	.95	

Note. Correlations above the diagonal are correlations from the weight-development sample. Correlations below the diagonal are from the cross-validation sample. All correlations are significant at $p < .001$. All statistics presented in this table are *n*-weighted means of results from 61,180 students in 101 schools.

a. First value is for weight-development sample. Values in parentheses are for the cross-validation sample.

Table 4.

Variance Explained for FGPA-Derived Composites and Incremental Validity over HSGPA and SAT – *N*-Weighted by School

	Variance Explained for Composite[a]	% of Maximum Variance Explained[a]	Incremental Variance Explained over HSGPA and SAT[a]	% of Maximum Incremental Variance[a]	% Improvement over HSGPA and SAT[a]
1. Regression – Full 395 items	.096 (.078)		.010 (.007)		4.1% (2.8%)
2. ARI Composite	.073 (.069)	75.9% (89.2%)	.005 (.005)	45.7% (80.9%)	1.9% (2.3%)
3. Regression – 5 Domain Factors	.076 (.073)	78.9% (93.8%)	.005 (.006)	49.7% (90.8%)	2.1% (2.6%)
4. Regression – 25 points	.076 (.074)	79.2% (95.3%)	.005 (.006)	52.5% (95.4%)	2.2% (2.7%)
5. Regression – Best 25 Items	.076 (.068)	78.8% (87.2%)	.006 (.005)	56.3% (76.1%)	2.3% (2.2%)
6. Vertical Percent Method – 5 Domain Factors	.072 (.069)	75.2% (88.6%)	.005 (.006)	47.2% (88.6%)	2% (2.5%)
7. Vertical Percent Method – 25 points	.076 (.073)	78.9% (93.7%)	.005 (.006)	47.2% (86.4%)	2% (2.5%)
8. Vertical Percent Method – Full 395 Items	.059 (.053)	61.5% (68.5%)	.005 (.004)	51.9% (65.4%)	2.1% (1.9%)

Note. Variance explained estimates are *n*-weighted means of results from 67,644 students in 110 schools. Percentage estimates are simple percentages of the *n*-weighted value, and not *n*-weighted themselves. Percent of maximum variance explained and percent of maximum incremental variance both are relative to the full 395-item regression model. Percent improvement over HSGPA and SAT represents the additional variance explained after adding the composite to a model as a percentage of the variance already explained by HSGPA and SAT.

a. First value is for weight-development sample. Values in parentheses are for the cross-validation sample.

Table 5.

Variance Explained for Retention-Derived Composites and Incremental Validity over HSGPA and SAT – *N*-Weighted by School

	Variance Explained for Composite[a]	% of Maximum Variance Explained[a]	Incremental Variance Explained over HSGPA and SAT[a]	% of Maximum Incremental Variance[a]	% Improvement over HSGPA and SAT[a]
1. Regression – Full 395 items	.039 (.023)		.025 (.016)		42.6% (26.7%)
2. ARI Composite	.019 (.021)	50.2% (92.0%)	.009 (.014)	36% (85.5%)	21.1% (23.8%)
3. Regression – 5 Domain Factors	.020 (.021)	51.5% (93.8%)	.009 (.014)	37.3% (87.8%)	21.7% (24.3%)
4. Regression – 25 points	.022 (.022)	55.9% (98.2%)	.010 (.013)	40.9% (84.6%)	23.3% (23.6%)
5. Regression – Best 25 Items	.024 (.022)	60.9% (94.0%)	.013 (.015)	51.2% (92.5%)	27.5% (25.3%)
6. Vertical Percent Method – 5 Domain Factors	.019 (.021)	50% (90.3%)	.009 (.013)	35.6% (82.0%)	20.9% (23.1%)
7. Vertical Percent Method – 25 points	.021 (.022)	53.3% (94.1%)	.009 (.014)	38.1% (85.0%)	22.0% (23.7%)
8. Vertical Percent Method – Full 395 Items	.017 (.020)	44.1% (89.4%)	.008 (.016)	31.8% (100%)	19.1% (26.8%)

Note. Variance explained estimates are *n*-weighted means of results from 61,180 students in 101 schools. Percentage estimates are simple percentages of the *n*-weighted value, and not *n*-weighted themselves. Percent of maximum variance explained and percent of maximum incremental variance both are relative to the full 395-item logistic regression model. Percent improvement over HSGPA and SAT represents the additional variance explained after adding the composite to a model as a percentage of the variance already explained by HSGPA and SAT.

a. First value is for weight-development sample. Values in parentheses are for the cross-validation sample.

Appendix

Table A1.

Summary of Comparisons of Different Methods for Predicting FGPA and Retention – Total Sample

	r with FGPA	R^2 with Retention	Incremental R^2 for FGPA	Incremental R^2 for Retention
ARI Composite	.338 (.329)	.056 (.049)	.000 (.000)	.006 (.004)
Regression – all 395 items	.410 (.381)	.099 (.058)	.013 (.007)	.036 (.011)
Regression – 25 points	.357 (.350)	.066 (.055)	.002 (.002)	.011 (.007)
Regression – 5 domain factors	.351 (.343)	.061 (.051)	.001 (.001)	.008 (.004)

Note. Incremental variance explained is over HSGPA and SAT. For all cells, the first value is for the weight-development sample, whereas the value in parentheses is for the cross-validation sample. Statistics presented in this table are based on the total sample of results from 67,644 students in 110 schools for the FGPA criterion, and 61,180 students in 101 schools for the retention criterion.

Table A2.

Descriptive Statistics and Correlations Between Variables Used in Predicting FGPA – *N*-Weighted by School

	Mean[a]	SD[a]	1	2	3	4	5	6	7	8	9	10	11
1. FGPA	2.93 (2.93)	.74 (.73)		.44	.44	.34	.35	.36	.41	.37	.32	.34	.34
2. HSGPA	3.62 (3.61)	.50 (.51)	.43		.46	.47	.46	.46	.44	.44	.42	.48	.47
3. SAT	1662.7 (1660.7)	263.7 (261.7)	.43	.45		.63	.63	.63	.65	.63	.58	.63	.62
4. ARI Composite	.62 (.62)	.25 (.25)	.33	.46	.63		.96	.94	.80	.79	.77	.98	.94
5. Regression – 5 Domain Factors	.65 (.65)	.26 (.26)	.34	.45	.63	.96		.97	.83	.82	.79	.96	.93
6. Regression – 25 points	.48 (.48)	.26 (.26)	.35	.45	.62	.94	.97		.83	.82	.78	.96	.94
7. Regression – Full 395 Items	.37 (.37)	.30 (.30)	.38	.42	.64	.80	.83	.83		.91	.79	.80	.79
8. Regression – Best 25 Items	.20 (.20)	.27 (.27)	.36	.43	.61	.79	.82	.82	.91		.81	.81	.81
9. Vertical Percent Method – Full 395 Items	-2.58 (-2.56)	3.67 (3.68)	.31	.40	.58	.78	.79	.79	.79	.81		.80	.79
10. Vertical Percent Method – 25 points	-.15 (-.16)	2.76 (2.76)	.33	.47	.62	.98	.95	.96	.80	.81	.80		.96
11. Vertical Percent Method – 5 Domain Factors	-.02 (-.02)	.56 (.55)	.33	.46	.61	.94	.93	.94	.79	.81	.79	.96	

Note. Correlations above the diagonal are correlations from the weight-development sample. Correlations below the diagonal are from the cross-validation sample. All correlations are significant at $p < .001$. All statistics presented in this table are *n*-weighted means of results from 67,644 students in 110 schools.

a. First value is for weight-development sample. Values in parentheses are for the cross-validation sample.

Table A3.

Descriptive Statistics and Correlations Between Variables Used in Predicting Retention – Total Sample

	Mean[a]	SD[a]	1	2	3	4	5	6	7	8	9	10	11
1. HSGPA	3.62 (3.61)	.50 (.51)		.46	.19	.47	.45	.45	.36	.41	.41	.48	.47
2. SAT	1663.4(1662.1)	261.2 (259.6)	.45		.19	.62	.63	.61	.54	.60	.57	.62	.61
3. Retention	.87 (.86)	.34 (.34)	.19	.19		.18	.19	.19	.21	.20	.17	.18	.18
4. ARI Composite	1.37 (1.37)	.55 (.55)	.46	.62	.17		.96	.92	.68	.77	.76	.98	.94
5. Regression – 5 Domain Factors	1.48 (1.48)	.57 (.57)	.45	.63	.17	.96		.96	.71	.80	.78	.95	.93
6. Regression – 25 points	.95 (.95)	.59 (.59)	.44	.61	.18	.92	.96		.72	.82	.76	.94	.93
7. Regression – Full 395 Items	.76 (.75)	.85 (.87)	.35	.53	.16	.67	.70	.71		.80	.66	.69	.69
8. Regression – Best 25 Items	.62 (.62)	.67 (.67)	.41	.60	.17	.77	.80	.82	.78		.76	.78	.78
9. Vertical Percent Method – Full 395 Items	-1.35 (-1.33)	2.00 (2.01)	.39	.57	.16	.76	.78	.77	.65	.76		.78	.78
10. Vertical Percent Method–25 points	-.05 (-.05)	1.49 (1.49)	.47	.62	.17	.98	.95	.94	.68	.78	.79		.96
11. Vertical Percent Method – 5 Domain Factors	-.01 (-.01)	.30 (.30)	.46	.61	.18	.94	.93	.93	.67	.78	.78	.96	

Note. Correlations above the diagonal are correlations from the weight-development sample. Correlations below the diagonal are from the cross-validation sample. All correlations are significant at $p < .001$. All statistics presented in this table are based on a total sample size of 61,180 students in 101 schools.

a. First value is for weight-development sample. Values in parentheses are for the cross-validation sample.

Table A4.

Variance Explained for FGPA-Derived Composites and Incremental Validity over HSGPA and SAT – *N*-Weighted by School

	Variance Explained for Composite[a]	% of Maximum Variance Explained[a]	Incremental Variance Explained over HSGPA and SAT[a]	% of Maximum Incremental Variance[a]	% Improvement over HSGPA and SAT[a]
1. Regression – Full 395 items	.168 (.145)		.013 (.007)		4.6% (2.7%)
2. ARI Composite	.115 (.109)	68.2% (74.9%)	.000 (.000)	1.4% (1.4%)	0.1% (0.0%)
3. Regression – 5 Domain Factors	.123 (.117)	73.3% (81.0%)	.001 (.001)	7.3% (9.6%)	0.4% (0.3%)
4. Regression – 25 points	.128 (.122)	76.0% (84.3%)	.002 (.002)	13.5% (21.6%)	0.6% (0.6%)
5. Regression – Best 25 Items	.138 (.127)	82.4% (87.5%)	.005 (.004)	39.8% (51.0%)	1.9% (1.4%)
6. Vertical Percent Method – 5 Domain Factors	.115 (.109)	68.3% (75.1%)	.000 (.000)	2.0% (2.7%)	0.1% (0.1%)
7. Vertical Percent Method – 25 points	.117 (.111)	69.4% (76.9%)	.000 (.000)	1.4% (2.5%)	0.1% (0.1%)
8. Vertical Percent Method – Full 395 Items	.105 (.100)	62.3% (66.0%)	.001 (.001)	8.0% (8.1%)	0.4% (0.2%)

Note. Variance explained estimates based on a total sample size of 67,644. Percent of maximum variance explained and percent of maximum incremental variance both are relative to the full 395-item regression model. Percent improvement over HSGPA and SAT represents the additional variance explained after adding the composite to a model as a percentage of the variance already explained by HSGPA and SAT.

a. First value is for weight-development sample. Values in parentheses are for the cross-validation sample.

Table A5.

Variance Explained for Retention-Derived Composites and Incremental Validity over HSGPA and SAT – Total Sample

	Variance Explained for Composite[a]	% of Maximum Variance Explained[a]	Incremental Variance Explained over HSGPA and SAT[a]	% of Maximum Incremental Variance[a]	% Improvement over HSGPA and SAT[a]
1. Regression – Full 395 items	.099 (.058)		.036 (.011)		31.9% (12.3%)
2. ARI Composite	.056 (.049)	56.6% (84.2%)	.006 (.004)	15.6% (33.0%)	6.8% (4.4%)
3. Regression – 5 Domain Factors	.061 (.051)	61.5% (87.3%)	.008 (.004)	22.5% (40.6%)	9.5% (5.4%)
4. Regression – 25 points	.066 (.055)	66.5% (94.2%)	.011 (.007)	30.8% (60.6%)	12.6% (7.8%)
5. Regression – Best 25 Items	.076 (.057)	77.1% (98.0%)	.019 (.009)	53.3% (83.8%)	19.9% (10.5%)
6. Vertical Percent Method – 5 Domain Factors	.058 (.051)	58.5% (88.1%)	.007 (.005)	18.8% (44.6%)	8.1% (5.9%)
7. Vertical Percent Method – 25 points	.058 (.051)	58.9% (87.1%)	.006 (.004)	17.1% (36.2%)	7.4% (4.8%)
8. Vertical Percent Method – Full 395 Items	.051 (.045)	52.0% (77.5%)	.007 (.005)	19.1% (46.5%)	8.2% (6.1%)

Note. Variance explained estimates are based on a total sample size of 61,180. Percent of maximum variance explained and percent of maximum incremental variance both are relative to the full 395-item logistic regression model. Percent improvement over HSGPA and SAT represents the additional variance explained after adding the composite to a model as a percentage of the variance already explained by HSGPA and SAT.

a. First value is for weight-development sample. Values in parentheses are for the cross-validation sample.

The Research department actively supports the College Board's mission by:

- Providing data-based solutions to important educational problems and questions
- Applying scientific procedures and research to inform our work
- Designing and evaluating improvements to current assessments and developing new assessments as well as educational tools to ensure the highest technical standards
- Analyzing and resolving critical issues for all programs, including AP®, SAT®, PSAT/NMSQT®
- Publishing findings and presenting our work at key scientific and education conferences
- Generating new knowledge and forward-thinking ideas with a highly trained and credentialed staff

Our work focuses on the following areas

Admission	Measurement
Alignment	Research
Evaluation	Trends
Fairness	Validity

Follow us online: research.collegeboard.org

CollegeBoard

CollegeBoard

12b-6832