American Psychological Association

Journal of Applied Psychology

0021-9010/19/$12.00

2019, Vol. 104, No. 6, 814–831
http://dx.doi.org/10.1037/apl0000382

Effects of Range Restriction and Criterion Contamination on Differential Validity of the SAT by Race/Ethnicity and Sex

Jeffrey A. Dahlke, Paul R. Sackett, and Nathan R. Kuncel
University of Minnesota

We illustrate the effects of range restriction and a form of criterion contamination (individual differences in course-taking patterns) on the validity of SAT scores for predicting college academic performance. College data facilitate exploration of differential validity's determinants because they (a) permit the use multivariate range-restriction corrections to more accurately account for differential range restriction across subgroups and (b) allow for separate examinations of composite performance and specific performance episodes, the latter of which controls for ecological contamination of composite performance due to individuals' choices of performance opportunities. Using data from 363,004 students at 107 U.S. institutions, we found that controlling for course-taking patterns resulted in validity coefficients that were appreciably larger than predictors' correlations with obtained grade point averages (GPAs). The validities of SAT scores for predicting the first-year college performance of Black and Hispanic students were not significantly different from the validity for White students after correcting for both course-taking patterns and differential range restriction, but significant Black–White differences were detected for predicting 4-year cumulative performance. Validity estimates for predicting both first-year and 4-year cumulative performance were significantly smaller among Asian students than White students after making these corrections. The SAT's observed validity for predicting college GPAs was substantially lower for males than females and, unexpectedly, controlling for course-taking patterns increased male-female validity differences. Implications for personnel selection research are discussed.

Keywords: differential validity, range restriction, cognitive ability, standardized testing, criterion contamination

A common concern in personnel and educational selection research is that predictor variables (especially cognitive measures) may be more valid indicators of performance for some groups than for others, a phenomenon known as differential validity (Linn, 1978). The topic of differential validity has experienced renewed interest among industrial-organizational (I-O) psychologists within the past decade (e.g., Berry, Clark, & McClure, 2011), particularly with regard to the extent that range restriction might account for observed racial/ethnic differences in validity (e.g., Berry, Cullen, & Meyer, 2014; Berry, Sackett, & Sund, 2013; Roth, Le, Oh, Van Iddekinge, & Robbins, 2017). However, a difficulty in studying the effects of subgroup-specific range-restriction corrections on validity differences is the inability of conventional univariate corrections for range restriction (i.e., those performed when the unrestricted standard deviation can only be estimated for the predictor under study) to accurately model the actual selection mechanism(s) that produced restriction of range. We explore whether multivariate range restriction and differences in course-taking patterns (a form of criterion contamination) might account for the differential validity of SAT scores among racial/ethnic groups, between males and females, and among intersections of ethnicity and sex. Thus, in this article we use "subgroup" to refer to a group of people who are demographically differentiable on the basis of self-reported race and/or sex. As differential validity patterns tend to be fairly consistent across academic, military, and employee samples (Berry et al., 2011), our ability to correct for subgroup-specific multivariate range restriction and control for certain sources of criterion contamination in this study's academic database suggest fruitful directions for future personnel-selection research.

At the outset, we believe it is useful to delineate this article's focus on validity differences from other types of differences commonly studied with respect to racial/ethnic groups or sex. Our focus here is on *validity differences*, which reflect differences in the within-group predictive efficacy of test scores across groups (e.g., Are the correlations between test scores and performance similar or different across groups when computed separately within two or more groups?). This is quite different from research on *subgroup mean differences*, in which the focus is on the extent to which two groups' distributions of scores are different from

This article was published Online First January 14, 2019.

Jeffrey A. Dahlke, Paul R. Sackett, and Nathan R. Kuncel, Department of Psychology, University of Minnesota.

This research was supported by a grant from the College Board to Paul R. Sackett and Nathan R. Kuncel. Paul R. Sackett has served as a consultant to the College Board. This relationship has been reviewed and managed by the University of Minnesota in accordance with its conflict of interest policies. This research is derived from data provided by the College Board. Copyright ©2006-2013 The College Board. www.collegeboard.com.

Correspondence concerning this article should be addressed to Jeffrey A. Dahlke, Department of Psychology, University of Minnesota, 75 East River Road, Minneapolis, MN 55455. E-mail: dahlk068@umn.edu

each other in terms of central tendency. Research on mean differences is focused on univariate differences between groups, whereas the differential validity phenomenon is not affected by mean differences by virtue of examining within-group bivariate correlations in which data are centered and standardized within groups prior to analysis. Differential validity research is also different from research on *differential prediction* (i.e., predictive bias), which focuses on differences among subgroups' unstandardized subgroup regression lines, with the predictor and criterion variables analyzed in their unstandardized metrics (including whatever mean differences may exist on these variables). Although differential validity, differential prediction, and subgroup mean differences all have implications for the fairness and legal defensibility of selection systems, each concept addresses fairness in a different way. Beyond their common ties to the idea of fairness in assessment and selection, one thing that the literatures on differential validity, differential prediction, and subgroup mean differences have in common is that most research on these phenomena examines cognitive test scores, and the present study joins this trend.

Standardized test scores are widely used as predictor variables in postsecondary admissions processes and have a rather long history of documented differential validity by sex and ethnicity. With regard to sex differences, subgroup validity analyses of the SAT have consistently revealed that SAT scores are more valid indicators of college academic performance for females than males. For example, Mattern, Patterson, Shaw, Kobrin, and Barbuti (2008) found that the SAT was a more valid predictor of first-year grade point averages (GPAs) for females than males; this has been found to generalize when sex comparisons are made separately within ethnic groups (Bridgeman, McCamley-Jenkins, & Ervin, 2000). Mattern et al.'s (2008) data were gathered from students entering college in 2006, but similar trends have also been found using 1980s SAT data (Ramist, Lewis, & McCamley, 1990; Ramist, Lewis, & McCamley-Jenkins, 1994). These findings are consistent with research on cognitive-ability tests in employment settings, where validity estimates appear to be slightly larger for females than for males (Rothstein & McDaniel, 1992). Additionally, the validity of cognitive tests for predicting job performance tends to be greater for the sex that is dominantly represented in an occupation (Rothstein & McDaniel, 1992) and one could argue that females are the dominant sex in higher education because they attend college at a higher rate than males (Ma, Pender, & Welch, 2016).

While sex differences in SAT validity have been rather consistent over time and across studies, the SAT's ethnic-group validity differences for predicting first-year GPA have fluctuated somewhat. Ramist, Lewis, and McCamley-Jenkins (1994) found that the SAT's validity for predicting first-year GPAs was greater for White students than for Black or Hispanic students, but slightly greater for Asian students than for White students. Berry, Sackett, and Sund (2013) found that SAT scores were more valid for predicting first-year GPAs among White students than among Asian, Black, or Hispanic students, but the difference between validities computed for White and Asian students was quite small. Mattern et al. (2008) identified trends similar to Berry et al.'s (2013) results, but Mattern et al. (2008) reported larger validity differences between White and Asian students. These patterns of validity differences, with larger validity coefficients generally estimated for Whites, are not unique to the SAT. Berry, Clark, and McClure's (2011) meta-analysis examined the validity of cognitive tests in educational, military, and employment settings, finding that cognitive-ability tests tend to be slightly less valid predictors of performance for Black, Hispanic, and Asian groups than for White groups. Berry et al.'s (2011) meta-analysis included very large samples over many decades, which provided a robust summary of observed validity difference, but they were unable to account for an important statistical artifact: range restriction.

Statistical Artifacts Relevant to Differential Validity

Although differences in the validity of cognitive tests may certainly be real, it is also possible that these differences are due to, or exacerbated by, statistical artifacts. We attend to two artifacts that can have biasing effects on predictors' correlations with college performance: range restriction and criterion contamination in the form of differential course-taking patterns, which results in noncomparable GPAs across students (and potentially across groups).

Differences in Range Restriction Across Groups

Organizations' preferences for selecting high-ability individuals means that, on average, selectees have higher ability and are more similar in ability than are applicants. This reduction in variability caused by selection means that correlations among variables will be attenuated in samples of selectees. Differences in range restriction across subgroups could contribute to differences in observed validity coefficients, as statistics computed for more severely restricted groups (i.e., groups with lower mean scores on a predictor variable) will be more attenuated than statistics computed for less-restricted groups. Simulations have shown that range restriction could account for Black–White validity differences if selection decisions were based solely on the predictor of interest (Roth, Le Van Iddekinge, Buster, Robbins, & Campion, 2014), whereas research using real employee, student, and military data has shown that Black–White validity differences can persist even after separately accounting for univariate range restriction in Black and White samples (Berry et al., 2014). While Roth, Le Van Iddekinge, Buster, Robbins, and Campion's (2014) single-predictor selection scenario is unlikely to occur in practice, Berry et al.'s (2014) research illustrates that differential validity may truly exist and is worthy of further study.

Estimating differential range restriction's contribution to differential validity is a complicated venture because the most common corrections for range restriction can only account for the range restriction observed with respect to a single variable. Such corrections are only optimal when the predictor of interest was solely and explicitly used in the actual selection process (i.e., when direct range restriction occurs), but this is rarely the case; instead, most range restriction occurs in an indirect fashion, with selection occurring on one or more variables other than (or in addition to) the predictor one seeks to validate. Despite advancements in indirect range-restriction corrections that allow one to assume that the range restriction affecting a criterion is fully mediated through the predictor for which one has an applicant variance estimate (Hunter, Schmidt, & Le, 2006), the availability of range-restriction information for only one variable precludes a full account of range restriction's effects on validity estimates.

If range restriction occurs indirectly (or some combination of directly and indirectly), a univariate correction will be inadequate to estimate unrestricted differences in validity. In such a case, the ideal correction to use is the multivariate range-restriction correction (Aitken, 1934; Lawley, 1943), but researchers seldom have enough information to employ this correction because it requires knowledge of unrestricted covariances among predictors in addition to the unrestricted variances of these predictors. The multivariate correction is the general case of all other range-correction formulas and has the distinct advantage of accounting for how range restriction in multiple variables impacts the relationships among any number of other variables. The closer one gets to including all selection variables in a multivariate correction, the more accurately one will be able to estimate unrestricted validity coefficients. Thus, performing a multivariate correction on data from each of several subgroups allows for researchers to obtain more accurate differential validity estimates than are achievable by other means.

Multivariate corrections are commonly applied in research on standardized tests, as extensive information is available for both applicants and selectees in testing contexts. Research using educational data has detected validity differences between males and females and among ethnic groups even after applying multivariate corrections (Berry et al., 2013; Mattern et al., 2008; Ramist et al., 1994), which supports the ethnic-group validity differences reported by Berry et al. (2011, 2014). Findings from previous studies in which researchers applied corrections for range restriction to subgroup data have been fairly consistent, so we do not anticipate that range restriction alone will account for validity differences in our database. We therefore consider range restriction in combination with problems of criteria.

Differential Validity and Criterion Contamination

Another potential explanation for differential validity is criterion contamination, which means that a criterion variable is influenced by systematic variance not relevant to the performance construct of interest. In higher education, one salient source of contamination is individual differences in course-taking patterns (Berry & Sackett, 2009; Berry et al., 2013). Differential course-taking means that different students' GPAs are comprised of grades from different sets of courses (e.g., due to choices of major, elective courses, etc.), which, in turn, means that GPAs do not convey comparable information across students and that these GPA composites are not directly comparable across individuals (and potentially also not comparable across groups). Whereas range restriction is a purely statistical problem that can be corrected using well-known equations (see Sackett & Yang, 2000 for a review), differential course-taking and the noncomparability of students' GPAs can only be overcome by holding course-taking constant across students.

One strategy to control for differential course-taking is to switch the criterion of interest from GPAs to individual course grades (ICGs) and estimate validity at the level of the individual course (Berry & Sackett, 2009; Berry et al., 2013; Ramist et al., 1990, 1994). For example, Berry and Sackett (2009) and Berry et al. (2013) computed validity estimates for SAT scores using both college GPAs and ICGs as criteria. Then, to control for the effect of course-taking choices on SAT–GPA correlations, Berry and colleagues stepped-up the average SAT–ICG correlation into composite correlations between SAT scores and hypothetical GPA-like criteria representing what would happen if all students took the same curriculum of courses. Berry and Sackett (2009) found meaningful increases in validity estimates after controlling for course-taking patterns. Berry et al. (2013) applied this same approach to estimate differential validity estimates using multivariate range-restriction corrections and found that differences in course-taking did not account for validity differences in GPA. In this article, we revisit these findings using a newer SAT data set that is substantially larger in terms of students, schools, and courses than the data set analyzed by Berry et al. (2013) to conduct a more powerful test of the noncomparability hypothesis. We describe the logic underlying Berry et al.'s (2013) approach in greater detail in the following section.

The Present Study

The present study expands on previous research in three key ways. First, we estimate the validity of SAT scores for predicting college performance using a newer and substantially larger data set than has been analyzed in the past; this is important for estimating differential validity, as differential validity estimates become more stable as the number of minority group members increases and considerable statistical power is necessary to detect these differences. Second, we explore the effects of artifacts on differential validity separately by ethnicity and sex as well as among intersections of ethnicity and sex. Third, we study differential validity for predicting both first-year performance and 4-year cumulative performance. Previous studies of differential validity of the SAT have focused on first-year college performance, but 4-year cumulative GPA is also an important criterion that reflects longer-term academic performance (Berry & Sackett, 2009). In this study, we account for both range restriction and differences in course-taking patterns as we evaluate the influence of artifacts on validity differences. Our approach to account for course-taking patterns is admittedly complex, so we offer a conceptual introduction next.

To set the scene for the logic of the methods we use to control for differences in course-taking patterns, consider a situation in which one is tasked with validating a predictor of college faculty performance, but there are differences in the composition of job tasks across different subgroups in one's sample. Specifically, one subgroup performs a faculty job in a setting that is 90% teaching and 10% research, while another subgroup performs in a setting that is 10% teaching and 90% research. Given the extreme difference in the composition of performance demands between these contexts, a finding that the predictor variable exhibits different magnitudes of validity for predicting composite performance between the two subgroups should not be immediately attributed to a flaw in the predictor. Perhaps the predictor is simply more relevant for one dimension of performance than the other; follow-up analyses would be necessary to explore that possibility. What we do in this article is the conceptual equivalent of validating separately against the teaching criterion and the research criterion and comparing the pooled common-criterion validity across subgroups. Taking this approach allows us to determine whether the predictor really functions differently across groups, or whether aspects of the composite performance criterion may be

causing the subgroup validities to appear discrepant because of confounding factors.

The logic of the above example formed the basis of the methods used by Berry and Sackett (2009) and Berry et al. (2013) and we apply those same methods in the present study to examine whether differential validity exists across demographic subgroups after restricting analyses to settings in which all individuals are performing in the same context. As noted earlier, our data come from the postsecondary education domain and there is a long history of observing differential validity in this setting. However, the bulk of prior research has used GPAs as criteria without accounting for factors that can contribute to the noncomparability of GPAs across students and across subgroups. The specific factors that give rise to this noncomparability are beyond the scope of this article (e.g., Do these differences occur because students from certain groups tend to take more courses from particular disciplines or with particular levels of difficulty?). Instead, we are interested in exploring whether validity differences observed with GPA as the criterion are attenuated when validity is estimated only in courses in which all subgroups being compared are actually represented and these validities are then pooled across courses. Pooling validity estimates across the courses that allow head-to-head subgroup comparisons cuts across all dimensions on which courses differ and, as all courses provide data from all groups being considered, differences between groups on the pooled validity estimates cannot be attributed to differences in course-taking patterns. This approach offers a clearer indication of whether SAT scores are really less potent predictors for members of certain subgroups than is possible with GPA criteria. If the validity differences observed with GPA criteria are not reflected in the pooled course-level validities, it would signal that the noncomparability of students' GPAs operates as a confounding factor in analyses of differential validity.

As described above, pooling validity estimates across contexts in which all subgroups of interest are represented helps to identify whether validity differs across these subgroups after controlling for extraneous factors. However, the course-level validity estimates will not be directly comparable in magnitude to the validities estimated using observed composite criteria (e.g., GPAs). We are interested in whether the magnitudes of validity differences are smaller after accounting for artifacts compared with the differences observed in SAT–GPA correlations, but to make these types of comparisons we need to also account for the differences in the overall magnitudes of the validity estimates, as composite criteria are more reliable than are their components and are therefore generally more predictable. All else equal, predictor scores should correlate more strongly with a composite variable than with the composite's individual components; the difference in magnitude between predictor-component and predictor-composite correlations affects how one interprets the magnitude of differences between subgroup correlations. For example, a raw .05 difference between two groups' correlations means something very different when the magnitude of the referent group's correlation is .50 than when it is .20. To facilitate comparisons between the magnitudes of subgroup SAT–GPA and SAT–course grade correlations, we use composite correlation formulas to step-up the magnitudes of the pooled course-level correlations into a metric that is comparable with the metric of SAT-GPA correlations (see Appendix A for technical information on our compositing procedures and Appendix B for a worked example).

We call the composites that result from this estimation process "common-curriculum GPAs" because they represent what GPAs would be if all students took a full load of courses that are typical of the courses from our database that permitted head-to-head subgroup comparisons. The common-curriculum GPA is an abstraction that can be used to test hypotheses that cannot be directly addressed using students' observed GPAs. The fundamental idea is that although composite performance criteria may function differently in statistical analyses across groups because of factors that make those composite criteria noncomparable across individuals and groups, analyzing the components of those composites (e.g., individual course grades from college classes) instead of the composite variables themselves (e.g., college GPAs) can allow researchers to eliminate certain competing explanations for the differential validity observed with composite criteria.

Method

Participants

Our data were collected by the College Board in cooperation with colleges and universities that consider SAT scores in their admissions decisions. Participants in our study were a total of 363,004 students who began enrollment at 107 U.S. colleges and universities between 2006 and 2009. SAT scores, high school GPAs, and complete 4-year records of college GPAs and course-grade data were available for all students in our database. A breakdown of our sample by ethnicity, sex, and sex-ethnicity intersections is arrayed in Table 1. The 107 institutions in our data set varied on a number of important characteristics, including public versus private control, selectivity, size of student body, and regional location. The database used in our study has been used in other published research (e.g., Beatty, Walmsley, Sackett, Kuncel, & Koch, 2015; Dahlke, Kostal, Sackett, & Kuncel, 2018; Higdem et al., 2016; Kostal, Kuncel, & Sackett, 2015; Kostal, Sackett, Kuncel, Walmsley, & Stemig, 2017; Shewach, Shen, Sackett, & Kuncel, 2017; Yu, Sackett, & Kuncel, 2016), but the present study is the first to use these data to examine differential validity.

Measures

SAT composite scores. The College Board provided SAT scores for students at colleges participating in their data-collection

Table 1
Demographic Breakdown of Total Sample by Sex and Ethnicity

Ethnicity	Sex		Total
	Male	Female	
White	113,402 (31.2%)	127,643 (35.2%)	248,171 (68.4%)
Black	7,496 (2.1%)	12,194 (3.4%)	20,100 (5.5%)
Hispanic	11,758 (3.2%)	16,155 (4.5%)	28,289 (7.8%)
Asian	17,474 (4.8%)	18,658 (5.1%)	36,613 (10.1%)
Total	165,542 (45.6%)	194,555 (53.6%)	363,004

Note. Total ethnicity and sex percentages do not sum to 100% because (a) students had the option to not disclose demographic information on the self-report demographic questionnaire that accompanied the SAT and (b) very small minority groups (e.g., American Indian or Alaskan Native) are not tabled because there were not enough students from these groups to apply all of the analyses used in this study.

program. The SAT composite variable used in our substantive analyses is a simple sum of SAT Critical Reading and SAT Mathematics subtest scores. Scores on this composite could range from 400 to 1,600. College-applicant norms were available for a Critical Reading + Mathematics + Writing composite on which scores could range from 600 to 2,400, so we used this larger composite to make range-restriction corrections, as described in the Procedure section. We did not include writing scores in our substantive analyses because the SAT writing test had been made optional and was only required by a subset of schools at the time we began examining these data.

High school GPAs (HSGPAs). The College Board provided students' self-reported HSGPAs, which were collected at the time students took the SAT. College-applicant norms were available for self-reported HSGPAs, so we used this variable to make range-restriction corrections, as described in the Procedure section.

Individual course grades (ICGs). Each college reported ICGs for all courses taken by students at the institution. These ICGs were accompanied by course-identifying information, including the year and term in which students took the course and the college-assigned alpha-numeric course code (e.g., CHEM 100).

College GPAs. Colleges provided students' noncumulative and cumulative GPAs at the conclusion of each academic year. We used first-year GPAs and 4-year cumulative GPAs as measures of overall academic college performance.

Procedure

All procedures described below were performed using data from complete samples of students (without conditioning on demographics) as well as samples of demographically similar students grouped by ethnicity, sex, and intersections of ethnicity and sex. Each aspect of our procedure was therefore performed for each of our 15 sample types (i.e., for overall samples, for samples representing each of four ethnic groups, for males and females, and for eight sex-ethnicity intersections).

College GPA analyses. For our college GPA analyses, we computed correlations among first-year GPAs, 4-year cumulative GPAs, SAT scores (both the two-test and three-test composites), self-reported HSGPAs, and students' unweighted mean grades achieved during each year of college. The correlations between GPAs and the two-test SAT composite were of substantive interest and the other correlations were computed to allow the application of multivariate range-restriction corrections.

Individual course grade analyses. For our analyses of ICGs, we defined a course as a unique instance of a class with a given course code that occurred during a given academic year and term. To gauge trends in the predictability of ICGs as students progressed through college, we subdivided each course into cohorts of students who were taking the course during their first, second, third, or fourth year of college. We computed correlations among ICGs, SAT scores (both the two-test and three-test composites), and self-reported HSGPA for each student cohort within each course. As with our GPA analyses, the correlations between ICGs and the two-test SAT composite were of substantive interest, with the other correlations computed to facilitate multivariate range-restriction corrections.

In addition to computing validity coefficients within each course, we used course-grade information to compute the intraclass correlation coefficient for the ICGs earned in each year of college. These intraclass correlation coefficients indicate the average correlation among ICGs earned by individual students and were necessary for this study because we required an estimate of the average intercorrelation among ICGs to use in estimating common-curriculum GPAs, as described momentarily.

Corrections for range restriction. We used a multivariate range-correction procedure (Aitken, 1934; Lawley, 1943) to estimate how large all of our correlation coefficients would have been if all applicants had enrolled in college. We used subgroup-specific norms from schools' applicant populations to correct for range restriction in SAT scores and HSGPAs in our ICG and college GPA analyses. When school-specific applicant norms were not available for a particular subgroup, we used pooled norm information from institutions with a similar level of selectivity to make corrections. In executing these corrections, we used the range-restricted and the unrestricted covariance information from HSGPAs and the three-test SAT composite (Critical Reading, Mathematics, and Writing) to correct the covariances among the criterion variables (i.e., GPAs, ICGs) and the two-test SAT composite (Critical Reading and Mathematics) for range restriction. By including both the two-test and three-test SAT composites in our covariance matrices and using the three-test SAT composite with HSGPAs to make range-restriction corrections, we were able to capitalize on the additional information contained within the three-test composite while also obtaining corrected results for the two-test composite that was of primary interest. Including the three-test composite in our corrections was statistically equivalent to separately including each of the three SAT subtests as correction variables; this helped us to obtain more accurate corrected correlations, as incorporating more information into a range-restriction correction procedure leads to less biased corrected estimates. Range-restriction corrections were computed separately for each sample (i.e., individual corrections were made within each school or course).

Common-curriculum GPAs. To control for differential course-taking, we algebraically constructed composite variables representing what college GPAs would have been if all students had taken exactly the same set of courses throughout college, which we termed "common-curriculum GPAs." The process of computing the SAT's correlations with noncumulative common-curriculum GPAs is similar to how one might use the Spearman-Brown formula to estimate the reliability of a lengthened measure, but with the lengthening procedure applied to a validity coefficient instead of a reliability coefficient; in this case, ICGs are analogous to test items and GPAs are like composite test scores. Details of our computational procedure are outlined in Appendix A and a worked example is provided in Appendix B. We computed correlations between SAT scores and common-curriculum GPAs separately for each school and meta-analyzed those school-level estimates, as described in the following section.

Within each of our 15 demographic segments, we stepped-up the mean SAT–ICG correlation from each year of college at each school by the mean number of courses taken during the corresponding year of college at that school to estimate the correlation between SAT scores and a noncumulative common-curriculum GPA composite. In computing these stepped-up correlations, we used the intraclass correlation coefficients for ICGs to represent the average intercorrelation among ICGs earned during a particular

year of college. These noncumulative composite correlations represent how predictable noncumulative GPAs would be after controlling for the differentiating features of all of the courses the students took. Next, we aggregated these noncumulative common-curriculum GPA composites into a 4-year cumulative composite common-curriculum GPA using the formula for a weighted composite correlation, as described in Appendices A and B. In forming this cumulative composite criterion, we used the average correlation among mean grades across years of college at each school to account for year-to-year shared variance in the noncumulative common-curriculum GPAs.[1]

For a typical composite correlation in which all components come from the same sample and same level of analysis, the amount of sampling error in the estimate is indexed by the sample size of the components (see Schmidt & Hunter, 2015, p. 441). Our common-curriculum GPA composites, however, do not abide by this principle because the components came from different levels of analysis and were associated with different numbers of total observations. Some components were estimated at the school level (e.g., intercorrelations among mean ICGs) and were based on as many observations as there were students, whereas others were computed at the course level (e.g., SAT–ICG correlations) and the estimates of these components were based on more observations than there were individual students because most students took more than one course per year. Thus, we encountered a statistical dilemma when computing sampling variances for our composite correlations and, more importantly, the differences between subgroup composite correlations. We chose to resolve this dilemma by defining the sample size as the number of unique students who contributed to the common-curriculum GPA composite. By using the number of unique individuals as the basis for the sample size rather than the number of individual course grades, we were able to compute sampling variances for the common-curriculum GPA effect sizes that best reflect the uncertainty in our estimates.

Meta-analyses. We used Schmidt and Hunter's (2015) random-effects meta-analytic method to average the observed statistics from all samples extracted from our data set with sample-size weights. Specifically, we used the "ma_generic" function from the *psychmeta* R package (Dahlke & Wiernik, 2018, 2017/2018) to compute our meta-analyses, as that function allows researchers to use precomputed sampling variances to meta-analyze any effect size; this function was chosen because the sampling variances of our multivariate range-restriction corrected correlations had to be estimated in an ad hoc fashion. When meta-analyzing corrected correlations, we used the standard artifact-correction practice of scaling up the sampling variances by the square of the correction factors applied to the effect sizes (Schmidt & Hunter, 2015, pp. 143–145). We used random-effects meta-analytic standard errors of the mean validity estimates to compute the standard errors and statistical tests for all subgroup validity differences.

Subgroup comparisons. To ensure that subgroups' validity estimates would be comparable with each other in terms of the context in which the criterion data were obtained, we constrained our analyses to data that allowed head-to-head comparisons, such that all levels of a given demographic variable were represented in all samples analyzed. In comparisons among ethnicities, we only used data from samples in which all four ethnic groups of interest were represented; in comparisons between males and females, we only used data from sample in which both were represented; and in comparisons among sex-ethnicity intersections, we only used data from samples in which all eight sex-by-ethnicity groups were represented.

Results

Table 2 shows subgroup correlations between SAT scores and both first-year and 4-year cumulative performance criteria (including observed GPAs and the synthetic common-curriculum GPAs). The full sets of meta-analytic results corresponding to the data in Table 1 are tabled in Appendix C. Minority–White validity differences are summarized in Table 3 and are also depicted in Figure 1 for ease of interpretation, where negative values indicate that the validity estimates for students from minority groups are smaller than the estimates for White students. Male–female validity differences are summarized in Table 4, where negative values indicate that males had smaller validity estimates.

Overall Effects of Artifacts on Validity Estimates

A general pattern that emerged from our analyses was the tendency for the SAT composite to correlate more strongly with common-curriculum GPAs than with observed GPAs (see Table 2). The magnitude of validity gains increased even further after correcting for range restriction. This is consistent with previous research using synthetic performance composites to evaluate validity (Berry & Sackett, 2009; Berry et al., 2013) and lends support to the notion that differential course-taking can have a practically meaningful impact on validity estimates.

Observed Correlations With GPAs

Table 3 shows that observed first-year GPA validity differences were significant for Black–White contrasts in male, female, and mixed-sex samples; for Hispanic–White contrasts in male, female, and mixed-sex samples; and for Asian–White contrasts in male and mixed-sex samples, but not in female samples. Observed 4-year cumulative GPA validity differences were only significant for the Black–White contrast among females and the Asian–White contrast among males.

In terms of validity differences by sex, Table 4 shows that observed first-year GPA validity estimates were significantly smaller for males across all four ethnic groups as well as in samples that included all ethnicities. Observed 4-year cumulative GPA validity differences were significant in combined-ethnicity samples and in White, Hispanic, and Asian samples, but not in Black samples.

Observed Correlations With Common-Curriculum GPAs

One of the key questions of interest in this article is whether stepped-up composite correlations exhibit less differential validity

[1] Using the year-to-year correlations among mean grades is trivially different from using year-to-year correlations among GPAs (i.e., weighted mean grades), so we used the former because it is more consistent with the formulation of our common-curriculum GPA composites.

Table 2
Correlations Between SAT Composite Scores and Academic Performance Criteria

	First-year performance			4-year cumulative performance		
	Sex			Sex		
Ethnicity	Male	Female	Overall	Male	Female	Overall
			Observed correlations with GPAs			
White	.34 (.0078)	.39 (.0095)	.34 (.0079)	.31 (.0092)	.36 (.0101)	.29 (.0092)
Black	.29 (.0138)	.33 (.0130)	.30 (.0103)	.30 (.0189)	.31 (.0134)	.27 (.0125)
Hispanic	.29 (.0152)	.35 (.0119)	.31 (.0106)	.28 (.0158)	.33 (.0117)	.27 (.0114)
Asian	.29 (.0125)	.36 (.0149)	.31 (.0124)	.27 (.0130)	.33 (.0133)	.27 (.0114)
Overall	.36 (.0060)	.42 (.0060)	.37 (.0059)	.34 (.0062)	.40 (.0060)	.33 (.0059)
			Observed correlations with common-curriculum GPAs			
White	.34 (.0113)	.44 (.0124)	.37 (.0112)	.32 (.0101)	.43 (.0119)	.35 (.0102)
Black	.37 (.0212)	.42 (.0174)	.38 (.0171)	.36 (.0244)	.39 (.0195)	.36 (.0192)
Hispanic	.36 (.0191)	.44 (.0164)	.39 (.0158)	.34 (.0186)	.42 (.0159)	.36 (.0158)
Asian	.30 (.0149)	.39 (.0172)	.33 (.0144)	.29 (.0131)	.40 (.0173)	.33 (.0141)
Overall	.38 (.0082)	.48 (.0085)	.41 (.0082)	.36 (.0071)	.48 (.0081)	.40 (.0074)
			Correlations with GPAs corrected for range restriction			
White	.43 (.0077)	.48 (.0083)	.43 (.0077)	.40 (.0074)	.45 (.0077)	.38 (.0072)
Black	.37 (.0121)	.47 (.0160)	.42 (.0112)	.37 (.0164)	.44 (.0143)	.39 (.0123)
Hispanic	.41 (.0152)	.47 (.0112)	.43 (.0095)	.40 (.0149)	.45 (.0106)	.40 (.0098)
Asian	.39 (.0146)	.45 (.0159)	.40 (.0136)	.36 (.0139)	.42 (.0133)	.37 (.0117)
Overall	.47 (.0067)	.52 (.0054)	.48 (.0061)	.45 (.0061)	.50 (.0047)	.45 (.0054)
			Correlations with common-curriculum GPAs corrected for range restriction			
White	.47 (.0114)	.56 (.0115)	.49 (.0108)	.43 (.0099)	.53 (.0107)	.46 (.0093)
Black	.46 (.0193)	.54 (.0210)	.48 (.0177)	.40 (.0235)	.47 (.0192)	.41 (.0185)
Hispanic	.47 (.0191)	.55 (.0163)	.49 (.0157)	.41 (.0199)	.50 (.0153)	.43 (.0160)
Asian	.41 (.0180)	.49 (.0182)	.43 (.0165)	.37 (.0149)	.47 (.0168)	.40 (.0141)
Overall	.51 (.0087)	.59 (.0074)	.54 (.0078)	.47 (.0071)	.57 (.0073)	.52 (.0070)

Note. Values in parentheses are standard errors of the correlations between SAT scores and criteria. All results for ethnic, sex, and intersectional subgroups were computed using only samples in which at least three members of each group were present to allow head-to-head comparisons among groups at a common level of specificity.

than SAT–GPA correlations. When compared with differential validity estimates featuring observed GPA as the criterion, the results for SAT's correlations with common-curriculum GPAs estimated from observed ICG data suggest that controlling for course-taking patterns can explain many of the differences in subgroup validities. In fact, none of the Black–White or Hispanic–White validity differences were significant for common-curriculum GPAs (see Table 3). However, we still observed significant Asian–White validity differences for first-year performance and controlling for course-taking patterns also produced a significant Asian–White difference for first-year performance among females that was not detected for observed GPAs.

The results for both first-year and 4-year cumulative validity differences revealed significantly smaller validities among males than females in White, Hispanic, and combined-ethnicity samples, but not in Black samples (see Table 4).

Correlations With GPAs Corrected for Range Restriction

Having discovered that controlling for course-taking patterns can have a detectable impact on differential-validity trends, we turn our attention to the effects of range-restriction corrections applied to GPA criteria. Similar to the effect of controlling for course-taking patterns, correcting for range restriction reduced the magnitudes of most Black–White and Hispanic–White validity differences: Of these, only the first-year GPA Black–White validity difference among males was significant after correcting for range restriction (see Table 3). Asian–White validity differences among males were significant for both first-year and 4-year cumulative performance.

After correcting SAT–GPA relationships for range restriction, all male–female validity differences were significant for both first-year and 4-year cumulative performance (see Table 4).

Correlations With Common-Curriculum GPAs Corrected for Range Restriction

Given that applying separate corrections for range restriction and differences in course-taking patterns resulted in diminished magnitudes of differential validity in many cases, our final question is, how do these corrections affect validity differences when applied simultaneously? The bottom portion of Table 3 shows that none of the Black–White or Hispanic–White validity differences for first-year performance were significant in male, female, and mixed-sex samples after correcting for both course-taking and range restriction. Hispanic–White validity differences were also not significant for 4-year cumulative performance, but Black–

Table 3
Minority–White Validity Differences for SAT Composite Scores' Correlations With Academic Performance Criteria

Focal-group ethnicity	First-year performance			4-year cumulative performance		
	Sex			Sex		
	Male	Female	Overall	Male	Female	Overall
Observed correlations with GPAs						
Black	−.05** (.0159)	−.06** (.0161)	−.04** (.0130)	−.02 (.0210)	−.05** (.0168)	−.01 (.0155)
Hispanic	−.05** (.0171)	−.04** (.0152)	−.03* (.0132)	−.03 (.0182)	−.03 (.0155)	−.01 (.0146)
Asian	−.05** (.0148)	−.03 (.0176)	−.03* (.0147)	−.04** (.0159)	−.03 (.0167)	−.02 (.0147)
Observed correlations with common-curriculum GPAs						
Black	.02 (.0240)	−.02 (.0214)	.01 (.0204)	.04 (.0264)	−.03 (.0228)	.00 (.0217)
Hispanic	.02 (.0221)	−.00 (.0206)	.02 (.0194)	.02 (.0212)	−.00 (.0199)	.01 (.0188)
Asian	−.04* (.0187)	−.05* (.0212)	−.04* (.0182)	−.03 (.0165)	−.03 (.0210)	−.02 (.0174)
Correlations with GPAs corrected for range restriction						
Black	−.07** (.0144)	−.01 (.0180)	−.01 (.0136)	−.03 (.0180)	−.00 (.0162)	.01 (.0143)
Hispanic	−.02 (.0170)	−.01 (.0140)	.00 (.0122)	−.00 (.0167)	.01 (.0131)	.02 (.0122)
Asian	−.05** (.0165)	−.03 (.0179)	−.03 (.0156)	−.04** (.0157)	−.03 (.0154)	−.02 (.0137)
Correlations with common-curriculum GPAs corrected for range restriction						
Black	−.02 (.0224)	−.02 (.0240)	−.01 (.0207)	−.03 (.0255)	−.06** (.0220)	−.05* (.0207)
Hispanic	−.00 (.0222)	−.01 (.0199)	−.00 (.0190)	−.02 (.0222)	−.03 (.0187)	−.02 (.0185)
Asian	−.06** (.0213)	−.07** (.0215)	−.07** (.0197)	−.06** (.0179)	−.06** (.0199)	−.06** (.0168)

Note. Values in parentheses are standard errors of the validity differences. All results for ethnic, sex, and intersectional subgroups were computed using only samples in which at least three members of each group were present to allow head-to-head comparisons among groups at a common level of specificity. Negative values indicate smaller validities for the members of the focal ethnicity group than for White students.
* $p < .05$. ** $p < .01$.

White validity differences in female and mixed-sex samples were significant for this cumulative criterion. All Asian–White validity differences were significant for first-year and 4-year cumulative performance.

As reported above for the range-restriction corrected GPA analyses, all male–female validity differences were significant for both first-year and 4-year cumulative performance after correcting for both range restriction and differential course-taking patterns (see Table 4).

Discussion

We used data from students at 107 U.S. postsecondary institutions to study the effects of differential course-taking and range restriction on the differential validity of SAT scores by ethnicity, by sex, and by intersections of ethnicity and sex. We adopted the methods developed by Berry and Sackett (2009) and Berry et al. (2013) and used them to examine differential validity in both first-year and 4-year cumulative college performance with intersectional minority–White contrasts. After correcting for range restriction and controlling for criterion contamination due to individual differences in course taking, we found that Black–White and Hispanic–White validity differences were substantially mitigated when predicting first-year performance, but Asian–White and male–female validity differences were not. When predicting 4-year cumulative performance with these corrections applied, Asian–White and male–female validity differences persisted, and Black–White validity differences emerged.

We found that Black–White and Hispanic–White validity differences for first-year performance tended to be smaller after controlling for differential course-taking and correcting for range restriction using subgroup-specific norms and, after applying both corrections simultaneously, all Black–White and Hispanic–White validity differences were nonsignificant. These results are contrary to Berry et al.'s (2013) finding that making such corrections increased the magnitudes of validity differences among ethnic groups. We did, however, find that making simultaneous range-restriction and course-taking corrections to SAT scores' correlations with 4-year cumulative performance produced significant Black–White validity differences that were not detected in the observed GPA data or when the corrections were applied separately. Our data differed from Berry et al.'s (2013) data in terms of the time period, number of students, and number of schools represented. Thus, differences in results could be due to demographic changes over time, differences between cohorts, and/or changes to applicant populations that affected the norms used to make range-restriction corrections; we do not have the information necessary to determine which explanation is the most likely. Consistent with previous research (e.g., Mattern et al., 2008; Ramist et al., 1990, 1994), we found that SAT scores were more valid predictors of college performance for females than for males. Interestingly, we found that these sex differences were made larger by controlling for differential course-taking. It appears that, on average, females' college grades are more predictable than males' grades, even after one accounts for range-restriction artifacts and differences in course-taking patterns.

We were able to make more thorough corrections for artifacts in our academic data than are typically possible in other contexts (e.g., personnel selection) because of the College Board's system-

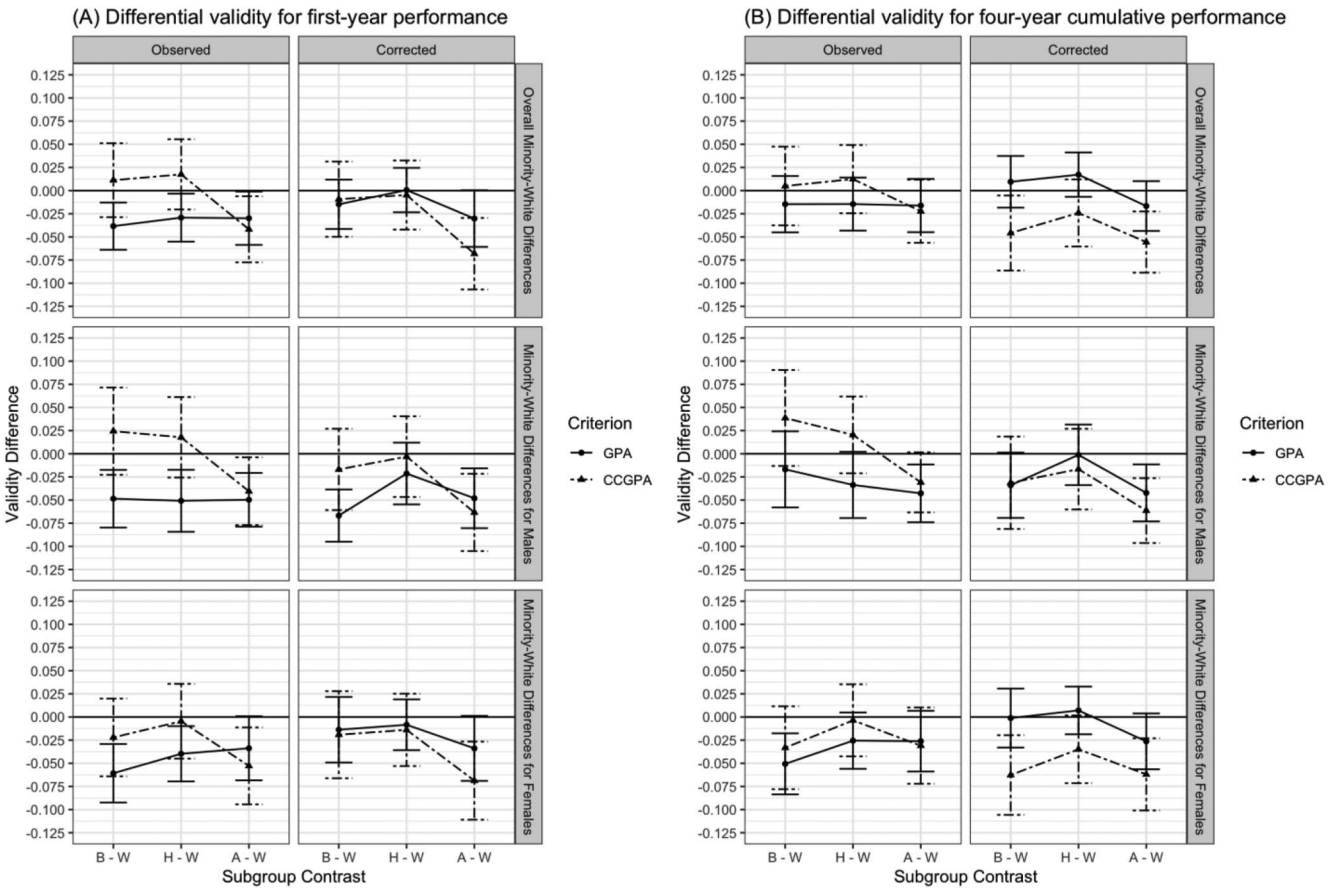

Figure 1. Validity differences by ethnicity for predicting first-year and 4-year cumulative performance from SAT composite scores (i.e., the sum of SAT Critical Reading and SAT Mathematics scores) with 95% confidence intervals. CCGPA = common-curriculum GPA; W = White; B = Black; H = Hispanic; A = Asian. A negative validity difference indicates that the validity for a minority group (i.e., Black, Hispanic, or Asian students) is smaller than for White students. Columns of the plot grids organize results by correlation type (observed correlations vs. correlations corrected for multivariate range restriction).

atic collection of applicant and enrolled-student predictor data and the availability of specific course-level and composite GPA indices of performance. Our ability to make corrections for multivariate range restriction meant that we could anchor our statistical corrections to two important predictors of performance (i.e., HSGPA and an SAT composite of critical reading, mathematics, and writing scores) rather than assuming that the entire effect of selection on validities could be represented by the variability of a composite of Critical Reading and Mathematics SAT scores alone. By accounting for the effect of selection on multiple predictors, our multivariate corrections are more likely to approximate the actual magnitudes of validity differences than are the univariate corrections typically applied in validation research.

Implications

Our corrections for range restriction and our controls for differential course-taking illustrate the importance of attending to artifacts and using data from specific performance events when a composite performance criterion is contaminated or otherwise not comparable across persons or groups. Although the academic context certainly differs from other contexts in which validity is closely monitored (e.g., personnel selection), academic data provide the opportunity to explore broadly relevant topics that would be difficult to study in other settings. Thus, while we caution against overgeneralizing from collegiate samples, we suspect that our analyses of academic data reveal useful insights into the broader phenomena of criterion contamination and differential validity.

It is notable that our analyses showed diminished differential validity after accounting for artifacts, whereas Berry et al.'s (2013) usage of these corrections did not lead to reduced magnitudes of differential validity estimates. As indicated earlier, the differences between our findings and Berry et al.'s (2013) could be a simple matter of statistical power from our use of a larger database or a substantive shift in the college population from the 1990s to the 2000s. Regardless of the source(s) of the differences in results, our findings signal the potential usefulness of analyzing validity estimates computed for individual components of performance in addition to those computed for composite performance metrics. Based on our results, analyzing criteria at a more specific level of

Table 4
Male–Female Validity Differences for SAT Composite Scores' Correlations With Academic Performance Criteria

Ethnicity	First-year performance	4-year cumulative performance
Observed correlations with GPAs		
White	−.05** (.0123)	−.04** (.0137)
Black	−.04* (.0190)	−.01 (.0232)
Hispanic	−.06** (.0193)	−.05** (.0197)
Asian	−.07** (.0195)	−.06** (.0186)
Overall	−.06** (.0085)	−.06** (.0086)
Observed correlations with common-curriculum GPAs		
White	−.10** (.0168)	−.11** (.0156)
Black	−.05 (.0275)	−.04 (.0312)
Hispanic	−.07** (.0252)	−.09** (.0245)
Asian	−.08** (.0227)	−.11** (.0217)
Overall	−.10** (.0118)	−.11** (.0107)
Correlations with GPAs corrected for range restriction		
White	−.05** (.0114)	−.04** (.0107)
Black	−.10** (.0201)	−.07** (.0217)
Hispanic	−.06** (.0189)	−.05** (.0183)
Asian	−.06** (.0215)	−.06** (.0192)
Overall	−.05** (.0086)	−.05** (.0077)
Correlations with common-curriculum GPAs corrected for range restriction		
White	−.09** (.0162)	−.10** (.0146)
Black	−.09** (.0285)	−.07* (.0303)
Hispanic	−.08** (.0251)	−.08** (.0251)
Asian	−.08** (.0256)	−.10** (.0224)
Overall	−.08** (.0114)	−.10** (.0102)

Note. Values in parentheses are standard errors of the validity differences. All results for ethnic, sex, and intersectional subgroups were computed using only samples in which at least three members of each group were present to allow head-to-head comparisons among groups at a common level of specificity. Negative values indicate smaller validities for males than for females.
* $p < .05$. ** $p < .01$.

analysis and pooling the results across those specific criteria may be a useful approach when composite criteria are not comparable across persons. This approach could easily be applied in any setting wherein the components of composite performance variables are available.

Performance episodes are often treated as composites, but our findings show that performance composites marked by noncomparability due to differences in course-taking choices function differently from composites in which all individuals are constrained to take the same set of parallel courses. Future research in employment contexts should attend to the possible effect of criterion contamination on minority–White validity differences and explore whether differential validity is smaller in work settings where all individuals who hold similar roles also do similar tasks than in settings where there is high variability in individuals' actual job tasks. This research would be especially valuable for understanding differential validity in settings where individuals who nominally hold the same type of job can be assigned to perform very different sets of tasks. To revisit the faculty example we presented earlier, not all academic jobs have the same composition of research, teaching, and service responsibilities. If one were to evaluate differential validity using some overall evaluation of performance without accounting for the differences in the types of tasks individuals perform, one could draw misinformed conclusions about the real extent of validity differences because those differences could very well exist purely because of noncomparability of performance scores.

Job contexts with high variability in the tasks assigned to workers are likely to be influenced by contamination that is similar in nature to differential course-taking patterns, which may have meaningful impacts on estimates of criterion-related validity and validity differences. We view the exploration of this phenomenon in work settings as a very exciting future direction for personnel research. For example, when researchers suspect that overall measures of performance are contaminated by differences in performance opportunities, we recommend that they attend to this contamination by validating predictors against separate dimension-level criterion measures. Those dimension-level validity estimates can then be combined into a composite correlation using weights that reflect the importance of each performance dimension to the organization, which avoids the problem of the dimension-level criteria receiving different weight in supervisors' judgmental performance composites for individual employees. Although the supervisors' composite judgments may still be the preferred data to consider in decision making, standardizing the weights assigned to different job activities across persons is preferable for validation research.

Even in the most extreme differential validity scenarios in Figure 1 and Tables 3 and 4, we still observed practically meaningful levels of validity for all groups, whether segmented by sex, ethnicity, or sex–ethnicity intersections. Given the rarity of organizations using strictly mechanical processes to select applicants, the magnitudes of differences reported here are unlikely to make a practically meaningful impact on the overall validity of holistic appraisals of applicants unless other aspects of the system exacerbate the validity differences. Most selection systems use multiple pieces of information and rely on human judgment to make offers to applicants, which may dilute the impact of differential validity for specific predictors on the validity of the system or may introduce new sources of validity differences that are unrelated to the differences attributable to test scores. Thus, while subgroup validities may differ for individual predictors, we do not have data to indicate whether the subgroup validity differences commonly observed for cognitive tests will meaningfully impact differential validity of overall selection systems. We encourage future research exploring how differential validity in one or more components of a system influences the differential validity of the system as a whole.

Limitations

Our data came from "SAT schools" and we are unable to relate our results to the differential validity of other common admissions tests (e.g., the ACT). Additionally, the extent to which the academic selection context corresponds to personnel selection or other selection scenarios is unclear. While criterion contamination and range restriction are relevant to many domains, it remains to be seen whether these findings can be replicated in other settings. Future research extending our findings could examine differences between use of composite performance indices (e.g., supervisor

evaluations, annual sales performance) and metrics of performance from more specific performance episodes (e.g., customer/client ratings, daily performance metrics) to conceptually replicate our analyses.

Despite our large sample of schools and students, our results are not based on random samples of schools or students. However, the diversity of institutions represented in our data set and the large numbers of students from different demographic groups provide support for the generalizability of our findings in the academic context.

This article was devoted to examining whether the noncomparability of criteria across persons and groups could contribute to differential validity as a form of criterion contamination. Our findings show this noncomparability can matter and these findings raises additional questions about which aspects of GPA noncomparability matter in analyses of differential validity. For example, do members of different groups take different numbers of courses from different disciplines or with different levels of difficulty? These questions are potentially interesting areas for future research and would require large-scale investigations that are beyond the scope of what we can offer as follow-up analyses in the present article. With our descriptive findings as a starting point, we encourage future research on when and why noncomparability of composite performance scores impacts substantive conclusions about subgroup validity differences in both school and work settings.

Conclusion

We found that college performance is considerably more predictable than is suggested by college GPA's correlations with SAT scores after accounting for range restriction and differences in course-taking patterns that contribute to the noncomparability of GPAs. Whereas previous research found that SAT scores were less valid predictors of college academic performance for Black and Hispanic students than for White students, we found that validity differences for predicting first-year college performance from SAT scores were not significant after controlling for differential course-taking and correcting for subgroup-specific range restriction. When predicting 4-year cumulative performance after making these corrections, however, we did detect significant White–Black validity differences. Overall, our results indicate that the effects of criterion contamination and differential range restriction are important for understanding differential validity. We emphasize the importance of attending to sources of artifactual variation that could be contributing to observed differences in validities among subgroups in future research on differential validity in school and work settings.

References

Aitken, A. C. (1934). Note on selection from a multivariate normal population. *Proceedings of the Edinburgh Mathematical Society, 4,* 106–110. http://dx.doi.org/10.1017/S0013091500008063

Beatty, A. S., Walmsley, P. T., Sackett, P. R., Kuncel, N. R., & Koch, A. J. (2015). The reliability of college grades. *Educational Measurement: Issues and Practice, 34,* 31–40. http://dx.doi.org/10.1111/emip.12096

Berry, C. M., Clark, M. A., & McClure, T. K. (2011). Racial/ethnic differences in the criterion-related validity of cognitive ability tests: A qualitative and quantitative review. *Journal of Applied Psychology, 96,* 881–906. http://dx.doi.org/10.1037/a0023222

Berry, C. M., Cullen, M. J., & Meyer, J. M. (2014). Racial/ethnic subgroup differences in cognitive ability test range restriction: Implications for differential validity. *Journal of Applied Psychology, 99,* 21–37. http://dx.doi.org/10.1037/a0034376

Berry, C. M., & Sackett, P. R. (2009). Individual differences in course choice result in underestimation of the validity of college admissions systems. *Psychological Science, 20,* 822–830. http://dx.doi.org/10.1111/j.1467-9280.2009.02368.x

Berry, C. M., Sackett, P. R., & Sund, A. (2013). The role of range restriction and criterion contamination in assessing differential validity by race/ethnicity. *Journal of Business and Psychology, 28,* 345–359. http://dx.doi.org/10.1007/s10869-012-9284-3

Bridgeman, B., McCamley-Jenkins, L., & Ervin, N. (2000). *Predictions of freshman grade-point average from the revised and recentered SAT® I: Reasoning test* (College Board Research Report No. 2000–1). Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2000.tb01824.x/abstract

Dahlke, J. A., Kostal, J. W., Sackett, P. R., & Kuncel, N. R. (2018). Changing abilities vs. changing tasks: Examining validity degradation with test scores and college performance criteria both assessed longitudinally. *Journal of Applied Psychology, 103,* 980–1000. http://dx.doi.org/10.1037/apl0000316

Dahlke, J. A., & Wiernik, B. M. (2018). psychmeta: An R package for psychometric meta-analysis. *Applied Psychological Measurement*. Advance online publication. http://dx.doi.org/10.1177/0146621618795933

Dahlke, J. A., & Wiernik, B. M. (2018). *psychmeta: Psychometric meta-analysis toolkit* (Version 2.1.7). Retrieved from https://CRAN.R-project.org/package=psychmeta (Original work published 2017)

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences*. San Francisco, CA: Freeman.

Higdem, J. L., Kostal, J. W., Kuncel, N. R., Sackett, P. R., Shen, W., Beatty, A. S., & Kiger, T. B. (2016). The role of socioeconomic status in SAT-freshman grade relationships across gender and racial subgroups. *Educational Measurement: Issues and Practice, 35,* 21–28. http://dx.doi.org/10.1111/emip.12103

Hunter, J. E., Schmidt, F. L., & Le, H. (2006). Implications of direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology, 91,* 594–612. http://dx.doi.org/10.1037/0021-9010.91.3.594

Kostal, J. W., Kuncel, N. R., & Sackett, P. R. (2015). Grade inflation marches on: Grade increases from the 1990s to 2000s. *Educational Measurement: Issues and Practice, 35,* 11–20. http://dx.doi.org/10.1111/emip.12077

Kostal, J. W., Sackett, P. R., Kuncel, N. R., Walmsley, P. T., & Stemig, M. S. (2017). Within-high-school versus across-high-school scaling of admissions assessments: Implications for validity and diversity effects. *Educational Measurement: Issues and Practice, 36,* 39–46. http://dx.doi.org/10.1111/emip.12134

Lawley, D. N. (1943). A note on Karl Pearson's selection formulae. *Proceedings of the Royal Society of Edinburgh. Section A: Mathematical and Physical Sciences, 62,* 28–30. https://www.cambridge.org/core/journals/proceedings-of-the-royal-society-of-edinburgh-section-a-mathematics/article/iva-note-on-karl-pearsons-selection-formulae/4A0B37E19E4FB5144E32EDE9A1C6EE99

Linn, R. L. (1978). Single-group validity, differential validity, and differential prediction. *Journal of Applied Psychology, 63,* 507–512. http://dx.doi.org/10.1037/0021-9010.63.4.507

Ma, J., Pender, M., & Welch, M. (2016). *Education pays 2016: The benefits of higher education for individuals and society (Trends in Higher Education)*. Retrieved from https://trends.collegeboard.org/sites/default/files/education-pays-2016-full-report.pdf

Mattern, K. D., Patterson, B. F., Shaw, E. J., Kobrin, J. L., & Barbuti, S. M. (2008). *Differential validity and prediction of the SAT®* (Research

Report No. 2008–4). College Board. Retrieved from http://eric.ed.gov/?id=ED562614

Mulaik, S. A. (2010). *Foundations of factor analysis*. Boca Raton, FL: CRC Press.

Ramist, L., Lewis, C., & McCamley, L. (1990). Implications of using freshman GPA as the criterion for the predictive validity of the SAT. In W. W. Willingham, W. H. Angoff, R. Morgan, & L. Ramist (Eds.), *Predicting college grades: An analysis of institutional trends over two decades* (pp. 253–288). Princeton, NJ: Educational Testing Service.

Ramist, L., Lewis, C., & McCamley-Jenkins, L. (1994). *Student group differences in predicting college grades: Sex, language, and ethnic groups* (College Board Research Report No. 93–1). Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.1994.tb01600.x/abstract

Roth, P. L., Le, H., OH, I.-S., Van Iddekinge, C. H., Buster, M. A., Robbins, S. B., & Campion, M. A. (2014). Differential validity for cognitive ability tests in employment and educational settings: Not much more than range restriction? *Journal of Applied Psychology, 99,* 1–20. http://dx.doi.org/10.1037/a0034377

Roth, P. L., Le, H., Oh, I.-S., Van Iddekinge, C. H., & Robbins, S. B. (2017). Who r u?: On the (in)accuracy of incumbent-based estimates of range restriction in criterion-related and differential validity research. *Journal of Applied Psychology, 102,* 802–828. http://dx.doi.org/10.1037/apl0000193

Rothstein, H. R., & McDaniel, M. A. (1992). Differential validity by sex in employment settings. *Journal of Business and Psychology, 7,* 45–62. http://dx.doi.org/10.1007/BF01014342

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology, 85,* 112–118. http://dx.doi.org/10.1037/0021-9010.85.1.112

Schmidt, F. L., & Hunter, J. E. (2015). *Methods of meta-analysis: Correcting error and bias in research findings* (3rd ed.). Thousand Oaks, CA: Sage. http://dx.doi.org/10.4135/9781483398105

Shewach, O. R., Shen, W., Sackett, P. R., & Kuncel, N. R. (2017). Differential prediction in the use of the SAT and high school grades in predicting college performance: Joint effects of race and language. *Educational Measurement: Issues and Practice, 36,* 46–57. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/emip.12150/full

Yu, M. C., Sackett, P. R., & Kuncel, N. R. (2016). Predicting college performance of homeschooled versus traditional students. *Educational Measurement: Issues and Practice, 35,* 31–39. http://dx.doi.org/10.1111/emip.12133

Appendix A

Details of the Computations Used to Estimate Common-Curriculum GPA Composites

To estimate the correlation between SAT scores and a stepped-up composite of course grades within a single year of college, we applied the scalar formula for a unit-weighted composite given by Ghiselli, Campbell, and Zedeck (1981, p. 163; Equations 7–12). This formula requires three pieces of information: (1) the average correlation between SAT scores and individual course grades (ICGs) during a particular year of college; (2) the mean number of courses taken by students during that year of college; and (3) the average intercorrelation among course grades earned by an individual student during that year of college (estimated as an intraclass correlation coefficient). To compute the correlation between SAT scores and the common-curriculum GPA (CCGPA) associated with the *i*th year of college, we used the formula:

$$\hat{r}_{SAT,CCGPA_i} = \frac{\bar{r}_{SAT,\,ICG_i}\bar{k}_{Courses_i}}{\sqrt{\bar{k}_{Courses_i} + \bar{k}_{Courses_i}\left(\bar{k}_{Courses_i} - 1\right)\overline{ICC}_{ICG_i}}} \tag{A1}$$

where $\bar{r}_{SAT,\,ICG_i}$ is the mean validity of SAT scores for predicting individual course grades in the *i*th year of college, $\bar{k}_{Courses_i}$ is the mean number of courses taken by individual students during the *i*th year, and $\overline{ICC}_{ICG_i}$ is the intraclass correlation coefficient for individual course grades in the *i*th year of college. We used this process to estimate the association between SAT scores and noncumulative CCGPAs in the first, second, third, and fourth years of college.

After computing each of the noncumulative SAT–CCGPA correlations, we aggregated data across years to estimate the association between SAT scores and 4-year cumulative CCGPAs. To accomplish this, we used the matrix formula for a weighted composite correlation to account for the shared variance in CCGPAs across 4 years of college. The matrix used to composite the noncumulative CCGPA correlations into cumulative CCGPA correlations was structured in the following way:

$$\mathbf{R} = \begin{bmatrix} 1 & \hat{r}_{SAT,\,CCGPA_1} & \hat{r}_{SAT,\,CCGPA_2} & \hat{r}_{SAT,\,CCGPA_3} & \hat{r}_{SAT,\,CCGPA_4} \\ \hat{r}_{SAT,\,CCGPA_1} & 1 & \bar{r}_{\overline{ICG_1},\overline{ICG_2}} & \bar{r}_{\overline{ICG_1},\overline{ICG_3}} & \bar{r}_{\overline{ICG_1},\overline{ICG_4}} \\ \hat{r}_{SAT,\,CCGPA_2} & \bar{r}_{\overline{ICG_1},\overline{ICG_2}} & 1 & \bar{r}_{\overline{ICG_2},\overline{ICG_3}} & \bar{r}_{\overline{ICG_2},\overline{ICG_4}} \\ \hat{r}_{SAT,\,CCGPA_3} & \bar{r}_{\overline{ICG_1},\overline{ICG_3}} & \bar{r}_{\overline{ICG_2},\overline{ICG_3}} & 1 & \bar{r}_{\overline{ICG_3},\overline{ICG_4}} \\ \hat{r}_{SAT,\,CCGPA_4} & \bar{r}_{\overline{ICG_1},\overline{ICG_4}} & \bar{r}_{\overline{ICG_2},\overline{ICG_4}} & \bar{r}_{\overline{ICG_3},\overline{ICG_4}} & 1 \end{bmatrix}$$

(*Appendices continue*)

where, in the lower triangle of the matrix, the first column contains the noncumulative SAT–CCGPA correlations estimated using Equation A1 and the second through fourth columns contain the correlations among students' mean grades in each year of college. On average, the variance of grades earned by a given student decreases from one year of college to the next (Dahlke et al., 2018). Thus, we accounted for variation in grade variance by constructing a diagonal scaling matrix **D** with standard deviations of criteria on the diagonal to use in computing our composite correlations.

$$\mathbf{D} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & \bar{s}_{ICG\ Year\ 1} & 0 & 0 & 0 \\ 0 & 0 & \bar{s}_{ICG\ Year\ 2} & 0 & 0 \\ 0 & 0 & 0 & \bar{s}_{ICG\ Year\ 3} & 0 \\ 0 & 0 & 0 & 0 & \bar{s}_{ICG\ Year\ 4} \end{bmatrix}$$

We defined a vector $\mathbf{w}_{CCGPA}$ to weight each year of criterion information using the mean number of courses taken by students during each of their first four years of college. A weight of zero was assigned to the SAT composite variable in this vector because this composite only aggregated criterion information.

$$\mathbf{w}_{CCGPA} = \begin{bmatrix} 0 & \bar{k}_{Courses_1} & \bar{k}_{Courses_2} & \bar{k}_{Courses_3} & \bar{k}_{Courses_4} \end{bmatrix}^T$$

To facilitate use of the matrix equation for composite correlations, we also constructed a vector for the predictor side of the equation in which SAT scores were given unit weight and all criterion variables were given weights of zero.

$$\mathbf{w}_{SAT} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \end{bmatrix}^T$$

With our correlation matrix, scaling matrix, and weight vectors prepared, we used the equation for a composite correlation (Mulaik, 2010, pp. 88–89) to estimate the correlation between SAT scores and 4-year cumulative CCGPAs:

$$\hat{r}_{SAT,Cumulative\ CCGPA} = \frac{\mathbf{w}_{CCGPA}^T\ \mathbf{D}\ \mathbf{R}\ \mathbf{D}\ \mathbf{w}_{SAT}}{\sqrt{\mathbf{w}_{CCGPA}^T\ \mathbf{D}\ \mathbf{R}\ \mathbf{D}\ \mathbf{w}_{CCGPA}}} \quad \text{(A2)}$$

where the numerator gives the covariance between SAT scores and cumulative CCGPAs and the denominator gives the standard deviation of the cumulative CCGPA composite. The variance of SAT scores was standardized, so it was not necessary to represent its standard deviation in the denominator. This method was applied to observed and corrected correlations, alike, and was used to estimate SAT–CCGPA correlations for each school so that the school-level validity estimates could be meta-analyzed.

Appendix B

A Worked Example of Common-Curriculum GPA Computations

As an additional aid for understanding our methodology and as an accompaniment to Appendix A, we have prepared a worked example of our common-curriculum GPA computations using mean statistics from our overall samples of students (i.e., students of all ethnicities and sexes). The process illustrated here was applied separately to all 15 of our demographic segments at each school. The first step in our process of estimating SAT scores' correlations with common-curriculum GPAs was to step-up the average correlation between SAT scores and individual course grades (ICGs) earned during a given year of college by the average number of classes taken in that year. In this example, the average SAT–ICG correlation for first-year coursework was .2495 and students took an average of 10.22 courses during their first year. To account for the dependency among the course grades earned by a typical student in this scenario, we used the intraclass correlation of .312 to represent the average intercorrelation among course grades in our composite. The composite formula indicates that, on average, SAT could be expected to correlate .405 with first-year common-curriculum GPAs.

$$\hat{r}_{SAT,CCGPA\ Year\ 1} = \frac{.2495 \times 10.22}{\sqrt{10.22 + 10.22 \times (10.22 - 1) \times .312}} = .405$$

Applying the procedure described above to all 4 years of noncumulative data indicated that SAT scores would correlate .405, .371, .322, and .258 with first-, second-, third-, and fourth-year noncumulative common-curriculum GPAs, respectively. We then organized these correlations in a matrix along with the average intercorrelations among the mean ICGs students earned in each year. For example, the average correlation between mean first-year ICGs and mean second-year ICGs is .675, as reflected in this matrix:

$$\mathbf{R} = \begin{bmatrix} 1 & .405 & .371 & .322 & .258 \\ .405 & 1 & .675 & .580 & .522 \\ .371 & .675 & 1 & .688 & .599 \\ .322 & .580 & .688 & 1 & .719 \\ .258 & .522 & .599 & .719 & 1 \end{bmatrix}$$

As the above correlation matrix is standardized, we created a scaling matrix in which the standard deviations of mean ICGs were arrayed on the diagonal. The standard deviations of mean ICGs for our overall sample in the first through fourth years were .785, .771, .757, and .754, respectively. SAT scores were left in *z*-score form, as only the criterion variances needed to be specified in our composite equations because the SAT was not being combined with anything.

(*Appendices continue*)

$$\mathbf{D} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & .775 & 0 & 0 & 0 \\ 0 & 0 & .771 & 0 & 0 \\ 0 & 0 & 0 & .757 & 0 \\ 0 & 0 & 0 & 0 & .754 \end{bmatrix}$$

After defining our correlation matrix and the diagonal scaling matrix, we created a vector of weights with which to combine our criterion variables. The mean weights assigned to first- through fourth-year noncumulative common-curriculum GPAs were 10.22, 10.05, 9.63, and 8.83, respectively. These weights represent the mean numbers of courses taken in each year.

$$\mathbf{w}_{CCGPA} = [0 \quad 10.22 \quad 10.05 \quad 9.63 \quad 8.83]^T$$

The other vector of weights we defined specified the weights give to predictors; as we only had one predictor, the SAT received a weight of 1 and the other variables received weights of 0.

$$\mathbf{w}_{SAT} = [1 \quad 0 \quad 0 \quad 0 \quad 0]^T$$

Finally, by using matrix multiplication, we computed the correlation between SAT scores and 4-year cumulative common-curriculum GPAs. In this case, that correlation was .40261.

$$\hat{r}_{SAT,Cumulative\ CCGPA} = \frac{\mathbf{w}_{CCGPA}^T \ \mathbf{D}\ \mathbf{R}\ \mathbf{D}\ \mathbf{w}_{SAT}}{\sqrt{\mathbf{w}_{CCGPA}^T \ \mathbf{D}\ \mathbf{R}\ \mathbf{D}\ \mathbf{w}_{CCGPA}}} = \frac{10.14472}{\sqrt{634.9084}} = .40261$$

Appendix C

Table C1
Meta-Analyses of Observed Correlations Between Composite SAT Scores and First-Year GPAs

Sex	Ethnicity	k	N	$\bar{r}$	$SE_{\bar{r}}$	SD_r	SD_{res}	95% CI	80% CV
All	All	107	363,004	.37	.0059	.06	.06	[.35, .38]	[.29, .44]
	White	46	196,192	.34	.0079	.05	.05	[.32, .35]	[.27, .40]
	Black	46	16,153	.30	.0103	.07	.05	[.28, .32]	[.23, .36]
	Hispanic	46	23,948	.31	.0106	.07	.06	[.29, .33]	[.23, .38]
	Asian	46	30,635	.31	.0124	.08	.08	[.28, .33]	[.21, .40]
Male	All	103	165,527	.36	.0060	.06	.06	[.34, .37]	[.28, .43]
	White	38	90,563	.34	.0078	.05	.04	[.33, .36]	[.29, .40]
	Black	38	5,629	.29	.0138	.09	.04	[.27, .32]	[.24, .35]
	Hispanic	38	9,873	.29	.0152	.09	.07	[.26, .32]	[.20, .39]
	Asian	38	14,271	.29	.0125	.08	.06	[.27, .32]	[.22, .37]
Female	All	103	194,304	.42	.0060	.06	.06	[.41, .43]	[.35, .49]
	White	38	99,387	.39	.0095	.06	.06	[.37, .41]	[.32, .46]
	Black	38	9,031	.33	.0130	.08	.06	[.31, .36]	[.26, .41]
	Hispanic	38	13,221	.35	.0119	.07	.06	[.33, .38]	[.28, .43]
	Asian	38	14,736	.36	.0149	.09	.08	[.33, .39]	[.26, .46]

Note. k = number of samples contributing to the meta-analysis; N = total sample size; $\bar{r}$ = weighted mean correlation; $SD_{\bar{r}}$ = standard error of mean correlation; SD_r = weighted standard deviation of correlations; SD_{res} = residual standard deviation of correlations; 95% CI = 95% confidence interval around $\bar{r}$; 80% CV = 80% credibility interval around $\bar{r}$. All *SD* and *SE* estimates were computed using unbiased formulas.

(*Appendices continue*)

Table C2
Meta-Analyses of Range-Restriction Corrected Correlations Between Composite SAT Scores and First-Year GPAs

Sex	Ethnicity	k	N	$\bar{r}$	$SE_{\bar{r}}$	SD_r	SD_{res}	95% CI	80% CV
All	All	107	363,004	.48	.0061	.06	.06	[.46, .49]	[.40, .55]
	White	46	196,192	.43	.0077	.05	.05	[.42, .45]	[.37, .49]
	Black	46	16,153	.42	.0112	.08	.03	[.39, .44]	[.38, .45]
	Hispanic	46	23,948	.43	.0095	.06	.03	[.41, .45]	[.39, .47]
	Asian	46	30,635	.40	.0136	.09	.08	[.37, .43]	[.30, .50]
Male	All	103	165,527	.47	.0067	.07	.06	[.46, .48]	[.39, .55]
	White	38	90,563	.43	.0077	.05	.04	[.42, .45]	[.38, .49]
	Black	38	5,629	.37	.0121	.07	.00	[.34, .39]	[.37, .37]
	Hispanic	38	9,873	.41	.0152	.09	.04	[.38, .44]	[.36, .46]
	Asian	38	14,271	.39	.0146	.09	.05	[.36, .41]	[.33, .44]
Female	All	103	194,304	.52	.0054	.05	.05	[.51, .53]	[.46, .59]
	White	38	99,387	.48	.0083	.05	.05	[.47, .50]	[.42, .54]
	Black	38	9,031	.47	.0160	.10	.05	[.44, .50]	[.40, .54]
	Hispanic	38	13,221	.47	.0112	.07	.03	[.45, .50]	[.44, .51]
	Asian	38	14,736	.45	.0159	.10	.08	[.42, .48]	[.35, .55]

Note. k = number of samples contributing to the meta-analysis; N = total sample size; $\bar{r}$ = weighted mean correlation; $SD_{\bar{r}}$ = standard error of mean correlation; SD_r = weighted standard deviation of correlations; SD_{res} = residual standard deviation of correlations; 95% CI = 95% confidence interval around $\bar{r}$; 80% CV = 80% credibility interval around $\bar{r}$. All *SD* and *SE* estimates were computed using unbiased formulas.

Table C3
Meta-Analyses of Observed Correlations Between Composite SAT Scores and First-Year Common-Curriculum GPAs

Sex	Ethnicity	k	N	$\bar{r}$	$SE_{\bar{r}}$	SD_r	SD_{res}	95% CI	80% CV
All	All	107	362,840	.41	.0082	.09	.08	[.39, .42]	[.30, .52]
	White	46	196,084	.37	.0112	.08	.07	[.35, .39]	[.27, .46]
	Black	46	16,148	.38	.0171	.12	.11	[.35, .41]	[.24, .51]
	Hispanic	46	23,932	.39	.0158	.11	.10	[.35, .42]	[.26, .51]
	Asian	46	30,612	.33	.0144	.10	.09	[.30, .35]	[.21, .44]
Male	All	103	165,420	.38	.0082	.08	.08	[.36, .39]	[.27, .48]
	White	38	90,485	.34	.0113	.07	.07	[.32, .37]	[.26, .43]
	Black	38	5,626	.37	.0212	.13	.11	[.33, .41]	[.23, .51]
	Hispanic	38	9,867	.36	.0191	.12	.10	[.32, .40]	[.23, .49]
	Asian	38	14,259	.30	.0149	.09	.08	[.27, .33]	[.20, .40]
Female	All	103	194,247	.48	.0085	.09	.08	[.46, .50]	[.37, .59]
	White	38	99,357	.44	.0124	.08	.07	[.42, .46]	[.34, .54]
	Black	38	9,029	.42	.0174	.11	.09	[.38, .45]	[.30, .53]
	Hispanic	38	13,211	.44	.0164	.10	.09	[.40, .47]	[.32, .55]
	Asian	38	14,725	.39	.0172	.11	.10	[.35, .42]	[.26, .51]

Note. k = number of samples contributing to the meta-analysis; N = total sample size; $\bar{r}$ = weighted mean correlation; $SD_{\bar{r}}$ = standard error of mean correlation; SD_r = weighted standard deviation of correlations; SD_{res} = residual standard deviation of correlations; 95% CI = 95% confidence interval around $\bar{r}$; 80% CV = 80% credibility interval around $\bar{r}$. All *SD* and *SE* estimates were computed using unbiased formulas.

(*Appendices continue*)

Table C4
Meta-Analyses of Range-Restriction Corrected Correlations Between Composite SAT Scores and First-Year Common-Curriculum GPAs

Sex	Ethnicity	k	N	$\bar{r}$	$SE_{\bar{r}}$	SD_r	SD_{res}	95% CI	80% CV
All	All	107	362,840	.54	.0078	.08	.08	[.53, .56]	[.44, .65]
	White	46	196,084	.49	.0108	.07	.07	[.47, .51]	[.40, .58]
	Black	46	16,148	.48	.0177	.12	.10	[.45, .52]	[.35, .62]
	Hispanic	46	23,932	.49	.0157	.11	.09	[.46, .52]	[.37, .61]
	Asian	46	30,612	.43	.0165	.11	.10	[.39, .46]	[.30, .56]
Male	All	103	165,420	.51	.0087	.09	.08	[.50, .53]	[.41, .62]
	White	38	90,485	.47	.0114	.07	.07	[.45, .49]	[.39, .56]
	Black	38	5,626	.46	.0193	.12	.06	[.42, .49]	[.38, .53]
	Hispanic	38	9,867	.47	.0191	.12	.09	[.43, .51]	[.35, .59]
	Asian	38	14,259	.41	.0180	.11	.09	[.37, .44]	[.30, .52]
Female	All	103	194,247	.59	.0074	.07	.07	[.58, .61]	[.50, .69]
	White	38	99,357	.56	.0115	.07	.07	[.54, .58]	[.47, .65]
	Black	38	9,029	.54	.0210	.13	.11	[.50, .58]	[.40, .68]
	Hispanic	38	13,211	.55	.0163	.10	.08	[.51, .58]	[.44, .65]
	Asian	38	14,725	.49	.0182	.11	.10	[.46, .53]	[.37, .61]

Note. k = number of samples contributing to the meta-analysis; N = total sample size; $\bar{r}$ = weighted mean correlation; $SD_{\bar{r}}$ = standard error of mean correlation; SD_r = weighted standard deviation of correlations; SD_{res} = residual standard deviation of correlations; 95% CI = 95% confidence interval around $\bar{r}$; 80% CV = 80% credibility interval around $\bar{r}$. All *SD* and *SE* estimates were computed using unbiased formulas.

Table C5
Meta-Analyses of Observed Correlations Between Composite SAT Scores and 4-Year Cumulative GPAs

Sex	Ethnicity	k	N	$\bar{r}$	$SE_{\bar{r}}$	SD_r	SD_{res}	95% CI	80% CV
All	All	107	363,004	.33	.0059	.06	.06	[.32, .35]	[.26, .41]
	White	46	196,192	.29	.0092	.06	.06	[.27, .31]	[.21, .37]
	Black	46	16,153	.27	.0125	.08	.07	[.25, .30]	[.18, .36]
	Hispanic	46	23,948	.27	.0114	.08	.07	[.25, .30]	[.19, .36]
	Asian	46	30,635	.27	.0114	.08	.07	[.25, .30]	[.18, .36]
Male	All	103	165,527	.34	.0062	.06	.06	[.33, .35]	[.26, .42]
	White	38	90,563	.31	.0092	.06	.05	[.30, .33]	[.24, .38]
	Black	38	5,629	.30	.0189	.12	.09	[.26, .33]	[.18, .41]
	Hispanic	38	9,873	.28	.0158	.10	.08	[.25, .31]	[.18, .38]
	Asian	38	14,271	.27	.0130	.08	.06	[.24, .30]	[.19, .35]
Female	All	103	194,304	.40	.0060	.06	.06	[.39, .41]	[.32, .47]
	White	38	99,387	.36	.0101	.06	.06	[.34, .38]	[.28, .43]
	Black	38	9,031	.31	.0134	.08	.06	[.28, .33]	[.23, .38]
	Hispanic	38	13,221	.33	.0117	.07	.06	[.31, .35]	[.26, .40]
	Asian	38	14,736	.33	.0133	.08	.07	[.30, .36]	[.24, .42]

Note. k = number of samples contributing to the meta-analysis; N = total sample size; $\bar{r}$ = weighted mean correlation; $SD_{\bar{r}}$ = standard error of mean correlation; SD_r = weighted standard deviation of correlations; SD_{res} = residual standard deviation of correlations; 95% CI = 95% confidence interval around $\bar{r}$; 80% CV = 80% credibility interval around $\bar{r}$. All *SD* and *SE* estimates were computed using unbiased formulas.

(*Appendices continue*)

Table C6
Meta-Analyses of Range-Restriction Corrected Correlations Between Composite SAT Scores and 4-Year Cumulative GPAs

Sex	Ethnicity	k	N	$\bar{r}$	$SE_{\bar{r}}$	SD_r	SD_{res}	95% CI	80% CV
All	All	107	363,004	.45	.0054	.06	.05	[.43, .46]	[.38, .51]
	White	46	196,192	.38	.0072	.05	.05	[.37, .40]	[.32, .44]
	Black	46	16,153	.39	.0123	.08	.03	[.37, .42]	[.35, .43]
	Hispanic	46	23,948	.40	.0098	.07	.03	[.38, .42]	[.36, .43]
	Asian	46	30,635	.37	.0117	.08	.06	[.34, .39]	[.29, .44]
Male	All	103	165,527	.45	.0061	.06	.05	[.44, .47]	[.38, .53]
	White	38	90,563	.40	.0074	.05	.04	[.39, .42]	[.35, .45]
	Black	38	5,629	.37	.0164	.10	.00	[.34, .40]	[.37, .37]
	Hispanic	38	9,873	.40	.0149	.09	.00	[.37, .43]	[.40, .40]
	Asian	38	14,271	.36	.0139	.09	.05	[.33, .39]	[.29, .43]
Female	All	103	194,304	.50	.0047	.05	.04	[.49, .51]	[.45, .55]
	White	38	99,387	.45	.0077	.05	.04	[.43, .46]	[.39, .50]
	Black	38	9,031	.44	.0143	.09	.01	[.42, .47]	[.43, .46]
	Hispanic	38	13,221	.45	.0106	.07	.01	[.43, .47]	[.44, .47]
	Asian	38	14,736	.42	.0133	.08	.06	[.39, .45]	[.34, .49]

Note. k = number of samples contributing to the meta-analysis; N = total sample size; $\bar{r}$ = weighted mean correlation; $SD_{\bar{r}}$ = standard error of mean correlation; SD_r = weighted standard deviation of correlations; SD_{res} = residual standard deviation of correlations; 95% CI = 95% confidence interval around $\bar{r}$; 80% CV = 80% credibility interval around $\bar{r}$. All *SD* and *SE* estimates were computed using unbiased formulas.

Table C7
Meta-Analyses of Observed Correlations Between Composite SAT Scores and 4-Year Cumulative Common-Curriculum GPAs

Sex	Ethnicity	k	N	$\bar{r}$	$SE_{\bar{r}}$	SD_r	SD_{res}	95% CI	80% CV
All	All	107	362,853	.40	.0074	.08	.08	[.39, .42]	[.30, .50]
	White	46	196,113	.35	.0102	.07	.07	[.33, .37]	[.27, .44]
	Black	46	16,146	.36	.0192	.13	.12	[.32, .39]	[.20, .51]
	Hispanic	46	23,920	.36	.0158	.11	.10	[.33, .40]	[.24, .49]
	Asian	46	30,612	.33	.0141	.10	.09	[.30, .36]	[.22, .44]
Male	All	103	165,442	.36	.0071	.07	.07	[.35, .38]	[.28, .45]
	White	38	90,515	.32	.0101	.06	.06	[.30, .34]	[.24, .40]
	Black	38	5,625	.36	.0244	.15	.13	[.31, .41]	[.19, .52]
	Hispanic	38	9,860	.34	.0186	.11	.10	[.30, .38]	[.21, .47]
	Asian	38	14,258	.29	.0131	.08	.07	[.26, .31]	[.20, .37]
Female	All	103	194,239	.48	.0081	.08	.08	[.46, .49]	[.38, .58]
	White	38	99,357	.43	.0119	.07	.07	[.40, .45]	[.34, .52]
	Black	38	9,028	.39	.0195	.12	.11	[.36, .43]	[.26, .53]
	Hispanic	38	13,207	.42	.0159	.10	.09	[.39, .46]	[.31, .54]
	Asian	38	14,726	.40	.0173	.11	.10	[.36, .43]	[.27, .52]

Note. k = number of samples contributing to the meta-analysis; N = total sample size; $\bar{r}$ = weighted mean correlation; $SD_{\bar{r}}$ = standard error of mean correlation; SD_r = weighted standard deviation of correlations; SD_{res} = residual standard deviation of correlations; 95% CI = 95% confidence interval around $\bar{r}$; 80% CV = 80% credibility interval around $\bar{r}$. All *SD* and *SE* estimates were computed using unbiased formulas.

(*Appendices continue*)

Table C8

Meta-Analyses of Range-Restriction Corrected Correlations Between Composite SAT Scores and 4-Year Cumulative Common-Curriculum GPAs

Sex	Ethnicity	k	N	$\bar{r}$	$SE_{\bar{r}}$	SD_r	SD_{res}	95% CI	80% CV
All	All	107	362,853	.52	.0070	.07	.07	[.50, .53]	[.43, .61]
	White	46	196,113	.46	.0093	.06	.06	[.44, .47]	[.38, .53]
	Black	46	16,146	.41	.0185	.13	.11	[.37, .45]	[.27, .55]
	Hispanic	46	23,920	.43	.0160	.11	.09	[.40, .46]	[.31, .55]
	Asian	46	30,612	.40	.0141	.10	.08	[.37, .43]	[.29, .51]
Male	All	103	165,442	.47	.0071	.07	.07	[.46, .49]	[.39, .56]
	White	38	90,515	.43	.0099	.06	.06	[.41, .45]	[.36, .50]
	Black	38	5,625	.40	.0235	.14	.11	[.35, .44]	[.26, .54]
	Hispanic	38	9,860	.41	.0199	.12	.00	[.37, .45]	[.41, .41]
	Asian	38	14,258	.37	.0149	.09	.07	[.34, .40]	[.28, .46]
Female	All	103	194,239	.57	.0073	.07	.07	[.56, .58]	[.48, .66]
	White	38	99,357	.53	.0107	.07	.06	[.51, .55]	[.45, .61]
	Black	38	9,028	.47	.0192	.12	.08	[.43, .51]	[.36, .58]
	Hispanic	38	13,207	.50	.0153	.09	.08	[.47, .53]	[.40, .60]
	Asian	38	14,726	.47	.0168	.10	.00	[.44, .50]	[.47, .47]

Note. k = number of samples contributing to the meta-analysis; N = total sample size; $\bar{r}$ = weighted mean correlation; $SD_{\bar{r}}$ = standard error of mean correlation; SD_r = weighted standard deviation of correlations; SD_{res} = residual standard deviation of correlations; 95% CI = 95% confidence interval around $\bar{r}$; 80% CV = 80% credibility interval around $\bar{r}$. All *SD* and *SE* estimates were computed using unbiased formulas.

Received October 23, 2017
Revision received November 15, 2018
Accepted November 15, 2018 ■