education policy analysis archives

A peer-reviewed, independent,
open access, multilingual journal

Arizona State University

Volume 23 Number 113 November 16th, 2015 ISSN 1068-2341

The Revised SAT Score and Its Potential Benefits for the Admission of Minority Students to Higher Education

Maria Veronica Santelices
Facultad de Educación, Pontificia Universidad Católica de Chile
Chile
&
Mark Wilson
University of California Berkeley
United States

Citation: Santelices, M. V., & Wilson, M. (2015). The revised SAT score and its potential benefits for the admission of minority students to higher education. *Education Policy Analysis Archives, 23*(113). http://dx.doi.org/10.14507/epaa.v23.2070

Abstract: This paper investigates the predictive validity of the Revised SAT (R-SAT) score, proposed by Freedle (2003) as an alternative to compensate minority students for the potential harm caused by the relationship between item difficulty and ethnic DIF observed in the SAT. The R-SAT score is the score minority students would have received if only the hardest questions from the test had been considered and was computed using a formula score and a regression approach. In this article we examine the potential effects of using the R-SAT of minority students in the admissions decision to selective institutions, and its capacity to predict short and long-term academic outcomes as well as its potential benefits regarding differential prediction of college grades for minority students. To test this out, we examined the performance of the R-SAT score compared to the standard SAT score in a sample of graduates from California public schools and in a subsample of

Journal website: http://epaa.asu.edu/ojs/
Facebook: /EPAAA
Twitter: @epaa_aape

Manuscript received: 24/03/2015
Revisions received: 13/09/2015
Accepted: 24/09/2015

students who enrolled in the University of California. We found that, in terms of the potential for college admissions for minority students, prediction power and the issue of overprediction, the R-SAT score did not perform significantly better than the SAT score.
Keywords: predictive validity; college admissions; SAT; revised SAT.

El Puntaje corregido del SAT y sus potenciales beneficios para la admisión de estudiantes minoritarios a la educación superior.
Resumen: En este trabajo se investiga la validez predictiva del puntaje corregido del SAT (R-SAT), propuesto por Freedle (2003) como una alternativa para compensar a los estudiantes de minorías étnica por el daño potencial causado por la relación entre la dificultad de los ítems y el funcionamiento diferencial del ítem (DIF) asociado a etnia observada en el SAT. El puntaje R-SAT es aquel que los estudiantes de minorías habrían recibido si se hubieran considerado solo las preguntas más difíciles de la prueba y se calcula utilizando una corrección por adivinación (*formula score*) y un enfoque de regresión. En este artículo se reflexiona sobre los efectos potenciales del uso de la R-SAT en la decisión de admisión a instituciones selectivas, su capacidad para predecir los resultados académicos de corto y largo plazo, y sus posibles beneficios en relación con la predicción diferencial de notas universitarias para estudiantes de minorías étnicas. Para estudiar esto, se analizó el desempeño de la puntuación R-SAT en comparación con la puntuación estándar SAT en una muestra de alumnos graduados de las escuelas públicas de California y en una submuestra de estudiantes que se matricularon en la Universidad de California. Los resultados muestran que la puntuación R-SAT no se comportó significativamente mejor que la puntuación SAT al considerar las posibilidades de admisión a la universidad de grupos minoritarios, el poder de predicción y el problema de sobrepredicción.
Palabras-clave: validez predictiva; admisión universitaria; SAT; SAT corregido.

A pontuação sat corrigida e o seu benefício potencial para estudantes de grupos minoritários no ensino superior.
Resumo: Este artigo pesquisa a validade preditiva da pontuação SAT Corrigida (R-SAT, pela sua sigla em inglês), proposta por Freedle (2003) como uma alternativa para balançar os grupos de estudantes minoritários de possíveis danos causados pelo vínculo entre a dificuldade e o funcionamento diferencial do item segundo a etnia (DIF, pela sua sigla em inglês). A pontuação R-SAT é a pontuação que os estudantes de grupos minoritários recebem se somente as perguntas mais difíceis do teste são consideradas e calculadas usando uma fórmula de pontuação e uma análise de regressão. Neste artigo, exploramos os efeitos potenciais do uso da R-SAT de estudantes de grupos minoritários nas decisões de admissão a instituições seletivas, e sua capacidade para predizer resultados académicos a curto e longo prazo, além de seus benefícios potenciais a respeito da predição diferencial das notas obtidas na faculdade dos estudantes de grupos minoritários. Para verificar isto na prática, examinamos o desempenho da pontuação R-SAT comparada com a pontuação do teste SAT padronizado numa amostra de estudantes formados das escolas públicas da Califórnia e uma sub-amostra de estudantes matriculados na Universidade da Califórnia. Os resultados mostram que em termos do potencial para a admissão no ensino superior de estudantes de grupos minoritários, o poder preditivo e o assunto de "superestimação", a pontuação R-SAT não é significativamente melhor do que a pontuação SAT.
Palavras-chave: validade preditiva; admissão à faculdade; SAT; SAT corrigida.

Introduction[1]

Admission examinations are often assessed by how well they predict college outcomes. Predictive validity studies analyze the degree of association between admissions test scores, like SAT scores, and college outcomes, such as college grades and graduation. These sorts of academic outcomes are relatively easy to collect and are also related to other important behaviors linked to success in college. Some studies have also addressed the ability of admission examination scores to predict nonacademic outcomes such as earnings, leadership, job satisfaction, satisfaction with life and civic participation (Allen, Robbins, & Sawyer, 2010; Bowen & Bok, 1998; Oswald, Schmitt, Kim, Ramsay, & Gillespie, 2004; Willingham, 1985).

In this study, we examine a measure of academic preparedness that has been proposed to complement the SAT. The "Revised-SAT" or R-SAT, was proposed by Roy Freedle (2003) with the goal of correcting, what he considered to be unfair results found through his application of the Standardization method for DIF, on the SAT results (Dorans & Holland, 1992; Dorans & Kulick, 1983, 1986). The R-SAT is based exclusively on a subset of the SAT questions—specifically, the more difficult items. There is a substantial body of literature on the validity of standardized test scores to predict college outcomes, however, we found no consensus among researchers about the predictive power of the SAT (Geiser & Santelices, 2007; Geiser & Studley, 2002; Ramist, Lewis, & McCamley-Jenkins, 1994; Zwick, 2002). The problem of differential prediction, or differential power of SAT scores for students from different ethnic groups on the prediction of college grades has also been extensively documented (Burton & Ramist, 2001; Geiser & Studley, 2002; Ramist et al., 1994; Zwick, Brown, & Sklar, 2004). In this context, the research presented in this article is important because it explores the potential benefits of considering the R-SAT for the admissions of minorities into higher education, especially into selective institutions, and compares it to the current use of the standard SAT score. If, on the one hand, Freedle´s findings and hypotheses about the R-SAT holds, the R-SAT would strengthen the validity of the use of test scores for admissions decisions of minority students and there would be room for debating about the most appropriate score to use for both White and minority students. On the other hand, a finding of little or no support for the R-SAT would weaken the arguments for the consideration of the R-SAT.

The Revised-SAT

Freedle observed a systematic relationship between item difficulty and differential item functioning in the SAT. This relationship is known as the "Freedle phenomenon": harder items were found to show DIF in favor of minority students while easier items tend to show DIF in favor of White students (Freedle, 2003). Differential item functioning (DIF) studies are used as the first step of fairness studies and refer to how items function after differences in score distributions between groups have been statistically removed. The remaining differences indicate that the items function differently for the two groups. Typically, the groups examined are derived from classifications such as gender, race, ethnicity, or socioeconomic status. The performance of the group of interest (focal group) on a given test item is compared to that of a reference or comparison group. White examinees are often used as the reference group, while minority students are often the focal groups (Holland & Wainer, 1993). DIF study´s results need to be complemented with the analysis of

[1] We wish to thank the College Board and the University of California Office of the President for providing the data. This research was supported, in part, by a UC ACCORD Dissertation Fellowship, and by Anillo Project SOC 1107 Statistics for Public Policy in Education from Conicyt. We appreciate the comments provided by David Stern, Robert MacCoun, Saul Geiser and Catherine Horn.

whether the source of difficulty difference is relevant or irrelevant to the test construct in order to judge the fairness for specific groups of students (Camilli, 2006).

Freedle proposed a new way to calculate the score to correct for the potential unfairness of the results caused by the systematic relationship between item difficulty and DIF observed in SAT items. The new score would capture how students perform on the hard half of the SAT test and is called the Revised-SAT or R-SAT (Freedle, 2003). The R-SAT would be provided to colleges as a complement to the SAT for minority students and would be the score that African American students would get if only hard questions were considered. Freedle described this score as a more valid assessment of African American`s knowledge. According to him, the R-SAT would increase the SAT verbal score by as much as 200 to 300 points for individual minority test-takers, it would reduce the mean score difference between White and minority test-takers by a third, and it would produce a score that is a better indicator of the academic ability of minority students.

Freedle, citing the work from Diaz-Guerrero and Szalay (1991), interprets the difference between a student's R-SAT and his/her regular SAT score as a measure of the degree to which the examinee's cultural background diverges from White, middle class culture. In his paper, Freedle recommends exploring the validity of the R-SAT index (a) by examining the correlation between the observed R-SAT index and college grades relative to the correlation between the observed SAT score and college grades and also, (b) by looking at how many admissions decisions would change if we use the R-SAT compered to using the SAT (where we would assume that, say, a score of over 600 indicates that a student qualifies for college). Freedle recognizes that such predictive validity analyses will necessarily be of limited interpretability because of the issue of restriction of range, as many of the students who would potentially be admitted by the R-SAT will be absent from the college grades data—nevertheless he considers it relevant to examine these predictions.

Freedle´s original work describing the relationship between item difficulty and DIF faced criticisms from several researchers (Camara & Sathy, 2004; Dorans, 2004, 2010; Dorans & Zeller, 2004a, 2004b; Wainer, 2009) despite the fact that others had already reported the phenomenon he observed (for example, Kulick & Hu, 1989; Schmidt & Bleistein, 1987).

His work was criticized by these researchers on technical grounds: (i) for the way Freedle had implemented the Standardization Approach to DIF and (ii) for using a dataset that preceded the ETS implementation of the bias and DIF sensitivity review for all items in the SAT. Freedle implemented the standardization approach using a non-standard denominator that did not consider omits and not-reached items and ignoring the fact that the SAT is a formula scored test. The sensitivity review was formalized at ETS in 1980. [2]

The official response from the College Board (Camara & Sathy, 2004) to Freedles' 2003 paper stressed the role of guessing in the phenomenon that Freedle described. This report blamed the systematic issue on students of low ability simply guessing the correct response to harder questions. This is also Bridgeman and Burton's contention (2005), which they illustrated using ad-hoc examples and the results from computerized testing. In addition, Bridgeman and Burton commented on the R-SAT, questioning its validity and reliability as an indicator of students' knowledge. Wainer (2009) also appealed to guessing to explain the phenomenon described by Freedle. He claim that the two parts used in the standardization methodology (stratification on total SAT and drawing inferences from a division of items into two parts: easy and hard) are contradictory if you consider that students can answer a particular item correctly not only based on ability but also based on chance. Central to the argument is the assumption that, on average, White

[2] In September 1980, ETS formalized a review process for test items that warranted the publication of the *ETS Test Sensitivity Review Process* (Ramsey, 1993). The process included several phases of qualitative review of both documents and tests. At that time, it did not include the analysis of DIF statistics.

students have higher ability level than African American students and that both groups have the same probability of guessing correctly. Under those assumptions, the observed relationship between item difficulty and DIF is to be expected, he says, due to a "statistical artifact".

Dorans (2004) and Dorans and Zeller (2004a) also criticized the methods Freedle used for calculating the necessary components of the R-SAT: the use of proportion correct rather than formula score, his consideration of different (ethnic) samples for the half-test and his application of inverse regression. Furthermore, Dorans and Zeller (2004b) explored the fairness of Freedle's R-SAT using Score Equity Assessment (SEA), a new methodology presented as a complement to the existing procedures for fairness assessment, namely DIF analysis and differential prediction. Using SEA Dorans and Zeller (2004b) found that the half-test to total test linking may be population-dependent and therefore the scores produced on the hard-half test cannot be used interchangeably with scores produced on the full-length SAT verbal test.

The Freedle phenomenon has been assessed with respect to a large new data set in a series of recent papers (Santelices & Wilson, 2010a, 2010b, 2012). The analyses reported in those papers use a modified R-SAT approach, incorporating the changes recommended by the ETS researchers, and the data sets all post-date the changes that were made to the ETS review procedures. Although the results show it to be less prevalent than Freedle originally reported, the existence of this phenomenon was found to be supported in general.

The research presented in this article sets aside the discussion about the "Freedle phenomenon," which has been largely centered on item functioning and its relationship with item difficulty, and focuses on the use of the R-SAT, highlighting its predictive validity. This paper examines the potential changes in admissions decisions for minorities if the R-SAT were used in combination with the SAT, its overall predictive validity and its potential benefit in differential prediction. In doing so, this study follows closely the recommendations made by both Freedle (2003) and his critic Dorans (2010). In the report written by Dorans (2010), he explicitly argues for the need to conduct predictive validity studies, not just DIF analysis, in order to address the questions raised by Freedle (2003):

> The fairness questions raised … about access to higher education are score-use questions that cannot be addressed by a DIF analysis … Differential prediction addresses score use. These studies typically assess whether test scores, alone or with other information such as high school grades, predict first-year grade point averages equally well for different subgroups (p. 2).

The Role of SAT Scores in the Prediction of College Outcomes

The argument for using standardized scores in admissions decisions, along with other indicators, relies heavily on their contribution to the prediction of college outcomes. Student-level variables such as motivation, academic performance and social integration have been identified by researchers as key factors in explaining college academic success (Bean & Mentzer, 1985; Pascarella & Terenzini, 1991; Tinto, 2006). There is a substantial body of literature on the power of standardized test scores in particular to predict a variety of college outcomes (Bowen & Bok, 1998; Camara & Echternacht, 2000; Kobrin, Patterson, Shaw, Mattern, & Barbuti, 2008; Willingham, 1985; Willingham, Lewis, Morgan, & Ramist, 1990) but we will focus our attention on the prediction of (i) college grades and (ii) graduation rates. Although these outcomes offer only a partial portrayal of student's educational achievement, the convenience of their collection and their frequent and systematic reporting makes them the outcomes most commonly used in predictive validity studies. Most often researchers have examined the predictive validity of standardized test scores and high school grades using short-term academic outcomes, especially grades. Long-term outcomes are often

assumed to be affected most significantly by financial aid and previous experience in college (Reason, 2009; Wilson, 1983) however the importance of graduation as the milestone and main reason for students pursuing higher education convinced us of the need to examine its prediction.

The findings from the literature are contentious: there is no consensus on the merits of the SAT to predict either short or long term outcomes (Geiser & Santelices, 2007; Geiser & Studley, 2002; Ramist et al., 1994; Zwick, 2002). Furthermore researchers have found that SAT scores do not predict equally well for students from different ethnic groups and, in particular, tend to overpredict the performance of Hispanics and African American students (Burton & Ramist, 2001; Geiser & Studley, 2002; Ramist et al., 1994; Zwick et al., 2004).

College Grade Point Average

The relationship between high school grade point average, SAT scores and freshmen grade point average has been widely examined by researchers at the College Board and research units within higher education institutions (e.g., Geiser & Studley, 2002; Ramist et al., 1994). In general the College Board studies find that SAT scores make a substantial contribution to predicting cumulative college GPAs and that the combination of SAT scores and high school records provide better predictions than either grades or test scores alone (Burton & Ramist, 2001; Hezlett et al., 2001). College Board researchers have studied the validity of the SAT mostly using correlational analysis and have taken into consideration the technical issues of range restriction, differences in grading across colleges and unreliability of college grades to measure success in college (Camara & Echternacht, 2000; Willingham et al., 1990). Typical correlations between first-year grades and the SAT I (Verbal and Math scores combined) range between 0.3 and 0.6 depending on the characteristics of the studies with an average of 0.4 (Ramist et al., 1994; Zwick, 2002). Bridgeman, Pollack, and Burton (2004) for example, report a correlation between freshman grades and the SAT I score composite of 0.55, while the SAT Verbal test score has a correlation of 0.50 with freshman grades, the SAT Math correlates 0.52.[3]

In 2005 the SAT I was revised in a number of ways (Kobrin, Patterson, Shaw, Mattern, & Barbuti, 2008) but still ETS researchers recommend the use of the test (especially the Writing test) in combination with high school grades when making admissions decisions since that combination maximizes predictability of first-year college grades (unadjusted r=0.46, r adjusted for range restriction=0.62). Non-ETS researcher and advocates, however, have stressed the low power of the SAT to predict college grades (FairTest, 2003; Geiser & Studley, 2002; Rothstein, 2004). Their arguments are based on the results of multivariate analyses that consider multiple academic predictors including student variables and school-level sociodemographics. For example, Geiser and Studley (2002), after taking the SAT II and high school GPA into consideration, reported that the SAT I scores improved the overall prediction rate by a negligible 0.1% (from 21.0% to 21.1%). The standardized coefficient of the SAT I, after controlling for SAT II and high school GPA, was 0.07, but statistically significant due, at least in part, to the large number of observations used.

Differential Prediction

Notable differences in the validity and predictive power of SAT scores and high school grades by race have been substantiated through numerous studies (e.g., Young, 2004). These two variables often overpredict the grades of African-American and Hispanic students and underpredict womens' performance (Burton & Ramist, 2001; Geiser & Studley, 2002; Ramist et al., 1994; Zwick et al., 2004). Overprediction means that a group's average predicted first-year grade point average

[3] They also report a correlation between high school grades and first-year college grades of 0.58.

(GPA) is greater than its average actual first-year GPA. Ramist, Lewis, and McCamley-Jenkins (1994) find that overprediction occurs even more strongly when using high school GPA alone (i.e., without the SAT) to predict first-year college grades.

Analyses of differential prediction are used to examine the bias of a test according to Cleary´s definition (1968), which defines bias against a specific subgroup as predictions of the criterion score obtained from a common regression line that are consistently too high or too low for members of that subgroup. There are a number of theories about the reasons for over and underprediction. Some have attributed the phenomenon to statistical artifacts (unreliability of the measures); others believed they are related to the differing college experiences of various student groups. Others hypothesize that students differ in ways that are not fully captured by either their test scores or high school grades. The observed facts still remain a matter of debate (Steele & Aronson, 1998; Zwick, 2002, 2006; Zwick et al., 2004).[4]

Researchers have looked at the differential prediction of test scores and high school grades among students from different language background (Zwick & Schemler, 2004; Zwick & Sklar, 2005) and from schools with different financial and teaching resources (Zwick & Himelfarb, 2011) as a way to investigate possible explanations to the issue of overprediction and underprediciton. Results show a reduction of prediction error for Hispanic and African American students, but not a complete elimination (from -0.15 to -0.08 and from -0.13 to -0.03 respectively) when using the second approach, and no change when considering first language.

College Graduation

From an economic perspective, the immediate goal of attending post-secondary education at the individual student level is college graduation (Hout, 2012). Studies exploring the role of SAT scores in college persistence and college graduation find only a moderate relationship (Astin, Tsui, & Avalos, 1996; Burton & Ramist, 2001; Mattern & Patterson, 2009, 2011a, 2011b). Wilson (1983) observes that the best predictor of college graduation are persistence to sophomore year and first-year GPA. This information is closest in time and in content to what is being predicted, and it is not available at admission. Studies attempting to predict interim persistence (return for sophomore year and five-semester persistence) have obtained low correlations, which range from 0.01 (high school grade point average) to 0.17 (SAT Math).

Although the traditional variables included in the multivariate regression models explain a small proportion of the variance associated to graduation, Geiser and Santelices (2007) found high school grades to be the strongest predictor, followed by the SAT II Writing scores. Zwick and Sklar (2005) corroborated the importance of high school grades. Sociodemographic variables play a minor role in explaining college graduation (Geiser & Santelices, 2007); nevertheless Bowen and Bok (1998) found these variables to be more important in the college prediction for African American students than for White students.

The lower correlation between college graduation and preadmission characteristics is to be expected since persistence in college and ultimate graduation are more substantially influenced by nonacademic factors than college GPA. Some of the non-academic variables that research has identified as playing an important role in determining persistence and graduation are finances,

[4]Although the stereotype threat theory (Steele & Aronson, 1998) has been suggested as a possible explanation to the over-prediction/underperformance phenomenon, it does not provide a straightforward account of the facts. If stereotype threat depressed standardized test performance, but did not affect subsequent academic work, we would expect to observe under-prediction (rather than over-prediction) of performance, because students would perform better in college than their test scores would indicate.

motivation, social adjustment, family and health problems, institution´s selectivity and size (Bowen, Chingos, & McPherson, 2009; Reason, 2009).

Non-Academic Predictors of College Success

A number of studies looking into the importance of non-academic variables to predict college success have claimed for the expansion of the definition of college success to include longer-term outcomes, such as persistence and graduation, as well as less-researched outcomes, such as leadership and civic participation (Camara & Kimmel, 2005; Kyllonen, 2008; Robbins, Lauver, Le, Davis, & Langley, 2004; Sternberg 1999, 2003). Doing so allows the prediction of college success using a broader range of indicators and thus avoiding the exclusive reliance on cognitive criteria and predictors. This seems a suitable recommendation in light of universities' broader missions, including social and personal outcomes for their students (Perfetto, 1999; Stemler, 2012) and the potential for reduced adverse impact on the admission of traditional minority students (Breland, Maxey, Gernard, Cumming, & Trapani, 2001; Oswald et al., 2004; Sinha, Oswald, Imus, & Schmitt, 2011; Sternberg, Gabora, & Bonney, 2012).

Why Are Standardized Tests Used in Admissions?

Despite the contentious arguments about the value of standardized tests in the prediction of college grades and graduation, higher education institutions continue to rely on standardized tests to make admissions decisions. Zwick (2002) justifies the use of the standardized test scores in admissions to large institutions by noting the cost of interviewing candidates or reviewing applications in elaborate detail. The cost for the school of collecting and processing the scores, she says, is very small compared to the cost of these alternatives. Tests allow all applicants the opportunity to perform in an environment with the same testing conditions, instructions and time-constraints. Standardized test scores allows the comparison of students who come from different schools in which grading standards can vary significantly.

Continued reliance of higher education institutions on standardized tests make alternative instruments and complementary scores especially relevant. The mixed conclusions from the research regarding the contribution of the SAT to the prediction of college grades and graduation, the overprediction of African American and Hispanic students´ performance in college, and the observed relationship between DIF and item difficulty, all call into question the validity of the use of SAT standardized test scores in admissions decisions. These validity issues should be considered in addition to the disparate effects of the SAT on minorities and their access to higher education.

Research Questions

The current paper explores the potential benefits of the R-SAT score for minority students. Rather than addressing the criticisms of design of the R-SAT (e.g., differential item functioning), we instead address the questions that bear on the use of the R-SAT, i.e., those that are most relevant for admissions officers:

1) Would use of the R-SAT score increase the number of minority students admitted to selective institutions?
2) Does the R-SAT score better predict the college outcomes of minority students than the SAT score?

3) Does the R-SAT score help ameliorate the issue of overprediction for African American and Hispanic students?

To answer these questions we first calculated the R-SAT and then we studied how beneficial it would be for minority students if the R-SAT were considered in admissions decisions at selective institutions. We compared the predictive power of the R-SAT, relative to the original SAT, both considering all students and then differentially by race. Finally, we analyzed whether the R-SAT score would help ameliorate the issue of overprediction for African American and Hispanic students. The predictive validity analyses considered the maximum score between the SAT Verbal score and R-SAT Verbal for minority students, not just the revised SAT score, in light of Freedle´s recommendation to report both scores and consider the difference between them as the extent to which there are cultural differences between White and minority students.[5] In addition, the regression models included sociodemographic variables based on the results of Rothstein (2004), who finds that most of the SAT predictive power comes from the correlation with sociodemographic variables. Although parental income and education play a modest role in the prediction of college performance when controlling for additional academic indicators such as high school grades and standardized tests (Geiser & Studley, 2002)[6], Rothstein's (2004) estimates show that the predictive contribution of the SAT I score is 60% lower than would be indicated by traditional methods that only consider academic variables.

Methodology

Data Sources

To investigate the first research question we drew from the College Board datafile of students from California public high school seniors who took the SAT forms DX and QI in 1994 or SAT forms IZ and VD in 1999 and spoke English as their best language. We only considered groups and forms in which the Freedle phenomenon has been observed and reported before (Santelices & Wilson, 2010a, 2010b, 2012). In particular, the R-SAT was calculated for African Americans in forms IZ, QI and DX and for Hispanics in forms IZ and VD (see Table 1).

Table 1
Number of Students for Whom the Revised Score Was Calculated

	1999 IZ	1999 VD	1994 QI	1994 DX	Total
White Examinees	6548	6682	3360	3188	19778
Hispanic Examinees	1904	2018	-	-	3922
African American Examinees	854	-	671	709	2234

[5] Predictive validity was also assessed using just the R-SAT and results do not provide stronger support for using the modified admission scores (see Appendix A).
[6] The authors reported an increase in R^2 from 22.3 to 22.8 when considering parental income and education in the regression equation that originally only included high school GPA, SAT I, and SAT II scores.

The College Board datafile allowed us to explore the research questions in a sample that is significant in size, especially for minority students, as it combines students from all public high schools in California. The College Board datafiles contained students' item level responses, and students´ individual scores, as well as students' responses to a Student Data Questionnaire (43 questions), which included self-reported demographic and academic information such as parents' education, family income, and high school grade point average. English as the best language is a standard requirement in DIF studies of the SAT similar to this as a way to analyze a group of students of common and mainstream educational experience and not confound DIF results with other educational needs (see Table 2).

In order to answer the second and third research questions, the information from the College Board just described was complemented with data from the University of California Corporate Data System which contains system wide admissions and performance data for all students who applied and then enrolled at UC. Through their applications to UC, students provide academic and demographic information that is subsequently verified and standardized. For those students who enroll at UC, this information contains their academic history as well—including college grades, number of courses, number of units completed and graduation. Information about parental education level and family income is also available for students who attended. An indicator of school performance on a state standardized test (Academic Performance Index) from the California Department of Education (2014) was also added to the file. The school academic performance index information was not available for the students who took the SAT in 1994 because the index was calculated for the first time in 1998, thus only results for students taking the SAT forms IZ and VD in 1999 are presented. This dataset allows us to explore the research questions in a sample that is significant in size, especially for minority students, as it combines students from nine University of California campuses.

As result of the eligibility criteria and of enrollment decisions, the sample used for the predictive validity analyses has a higher mean SAT score, higher high school grade point average, higher family income and parent's education than the College Board sample of all high school juniors from California public high schools who took SAT forms DX and QI in 1994 and SAT forms IZ and VD in 1999 and was used to answer the first research question (see Table 3).

Analyses

This section presents the details of how the R-SAT score was calculated and how the relative predictive power of these scores was assessed. Since previous studies found stronger evidence of the relationship between DIF estimates and item difficulty in the Verbal test than in the Mathematics test (Santelices & Wilson, 2010a, 2010b, 2012), all the analyses focus on the Verbal test although always controlling for the Mathematics scores.

The analyses exploring the impact of Freedle's R-SAT in admissions decisions and subsequent analyses looking at the R-SAT's predictive validity and differential prediction consider the maximum score between the SAT Verbal score and R-SAT Verbal score for minority students, and not just the revised SAT score. This is done in consideration of Freedle's own recommendations: "the solution is to recognize that this is pervasive phenomenon that can be easily remedied by reporting two scores, the usual SAT and the R-SAT" (Freedle, 2003). Since Freedle recommends reporting both scores for minority students and interprets the difference between them as an indication of the magnitude of the difference between the White majority's culture and the cultural background of minority groups, then the consideration of the maximum of the two scores

Table 2

Sample Descriptive Statistics by Ethnicity. College Bound Students from California Public High Schools who Took the SAT Forms DX and QI in 1994 or forms IZ and VD in 1999 and for Whom DIF was Observed

Variable	African American Students (Forms IZ, QI, DX)			Hispanic Students (Forms IZ, VD)			African American and Hispanic Students			White Students		
	Mean	SD	n	Mean	SD	n	Mean	SD	N	Mean	SD	n
Parental Education	15	2.8	597	13	3.4	1,513	13.8	3.3	2109	17	2.5	7712
Family Income	46,040	35,647	555	56,680	57,641	1,371	53,640	52,478	1925	95,980	84,218	6512
SAT Verbal Score	382	105.2	2237	472	96.4	3922	439	108.5	6156	502	106.6	19,778
R-SAT Verbal Score	407	88.9	2234	484	78.6	3922	456	90.3	6156	-	-	-

Table 3

Sample Descriptive Statistics by Ethnicity. Students who Took Forms IZ and VD in 1999 and for Whom DIF was Observed and who Enrolled at the University of California

Variable	African American Students			Hispanic Students			African American and Hispanic Students			White Students		
	Mean	SD	n	Mean	SD	n	Mean	SD	N	Mean	SD	n
Parental Education	15	2.9	208	13	3.7	666	14	3.33	874	17	2.49	2719
Family Income	51,415	38,390	185	54,815	55,983	605	54,019	52,393	790	10,0524	86,923	2278
SAT Verbal Score	498	84	221	520	90	685	515	89	906	587	80	2836
R-SAT Verbal Score	504	73	98	521	82	685	519	81	783	-	-	-

for minority students we believe represents the less disadvantageous scenario in which minority groups might compete for admission into selective colleges.[7]

Calculation of the revised SAT score. The R-SAT was obtained by calculating the corresponding formula score[8] in the hardest half of the test for all students who took each test form and then assigning African American/Hispanic students the total score obtained by White students who performed similarly in the hard half of that specific test form. Specifically, in order to obtain the revised score for African American/Hispanic students, first a linear regression was estimated only among the White students who took each form. The linear regression was used then to predict their SAT scores using the formula score obtained in the hard half of the test. A constant and a slope coefficient were estimated and subsequently those parameter estimates were applied to the formula score obtained, in the hard part of the test, by each African American and Hispanic student. This methodology, is the same as the one originally used by Freedle (2003), with the exception that we incorporated Dorans and Zeller's recommendations regarding the use of formula scores rather than the original proportion correct scores that Freedle used (Dorans, 2004; Dorans & Zeller, 2004a). The R-SAT thus allows one to estimate the number of correct responses (adjusted for random guessing) in a score metric that ranged from 200 to 800 just as for the regular SAT Verbal score. The scores of White students are used as the reference because they have been considered the reference group in previous DIF analyses.

Predictive validity analyses. The predictive power of the regular SAT verbal score and the R-SAT score were compared for African American, Hispanic, and White students. Linear regression was used for GPA prediction and logistic regression was used for the prediction of graduation (i.e., because UC GPA is a continuous numerical variable and graduation is a dichotomous outcome variable). Similar to the findings of Rothstein (2004), and contrary to the results reported by Bridgeman et al. (2004), visual inspection of scatterplot and the examination of linear, logarithmic, and exponential trends supported a linear relationship. The ordinary least squares method was used for estimating linear regressions and the maximum likelihood technique was implemented for the estimation of logistic regression. The college outcomes examined were the first through fourth year annual UC GPAs, the cumulative fourth year UC GPA, and whether students graduated by their fourth year at UC. All explanatory variables presented in models (1), (2), and (3) were introduced at once. No stepwise procedure was used. The academic outcomes included in this study are of particular interest because they are not limited to grade point averages and span four years of the college career of students taking the SAT in 1999.

Although sociodemographic covariates are not used in admissions, the analyses controlled for academic and sociodemographic variables found to be significant in previous college prediction research (Geiser & Studley, 2002; Zwick et al., 2004) because they change the estimated prediction power of test scores (Rothstein, 2004). The sociodemographic variables included parent's education and income level from the UC systemwide admissions and performance data. The academic

---

[7] The correlation between the SAT Verbal score and the maximum score between the SAT Verbal score and R-SAT Verbal score among minority students in the 1999 cohort is 0.948. Appendix A shows regression results using models that compare the predictive power of the original SAT Verbal score to that of the R-SAT score, in addition to those using the maximum score between the SAT and the R-SAT score. These models also exclude the API rank as explanatory variables. Results do not provide stronger support for using the modified admission scores.

[8] Formula scoring adjusts scores for the possibility of random guessing (Frary, 1988; Rogers, 1999).

variables included a weighted high school GPA, calculated with up to eight honors-level courses, the SAT Math score, and the school academic performance index expressed as quintile ranks for students who took the SAT in 1999.

Equations 1, 2 and 3 show the general regression equation models for the prediction of annual UC GPA, cumulative fourth-year UC GPA, and fourth-year UC graduation respectively.

$$UCGPA_{ikjs} = b_0 + b_1 APIQ_{ik} + b_2 Educ_{ik} + b_3 Inc_{ik} + b_4 HSGPA_{ik} + b_5 SATM_{ik} + b_6 Z_{iks} + e_{ikjs} \quad (1)$$

$$CUMUCGPA4_{iks} = q_0 + q_1 APIQ_{ik} + q_2 Educ_{ik} + q_3 Inc_{ik} + q_4 HSGPA_{ik} + q_5 SATM_{ik} + q_6 Z_{iks} + e_{iks} \quad (2)$$

$$LOGIT(GRAD4_{iks}) = a_0 + a_1 APIQ_{ik} + a_2 Educ_{ik} + a_3 Inc_{ik} + a_4 HSGPA_{ik} + a_5 SATM_{ik} + a_6 Z_{iks} \quad (3)$$

In model (1) $UCGPA_{ikjs}$ is the grade point average that a student *i*, of ethnicity *k* (where k can be equal to 1= African American, 2= Hispanic, 3= White) had in year *j* of college, considering verbal ability index *s*, where *j* ranges between 1 and 4 and *s* is either the SAT Verbal score (s=1) or the highest score between the R-SAT Verbal score and the original SAT score for minority students (s=2). $APIQ_{ik}$ refers to the ranking of the school attended by student *i* of ethnicity *k* in the California Academic Performance Index; $Educ_{ik}$ is the maximum number years of education achieved by the parents of student *i* of ethnicity *k* as reported in the UC application; Inc_{ik} refers to the family income of student *i* of ethnicity *k* reported in the UC application (expressed in dollars); $HSGPA_{ik}$ is the weighted high school GPA considering up to eight honors-level courses of student *i* of ethnicity *k* (which was the index used by UC at that time); $SAT\,M_{ik}$ is the score the student *i* of ethnicity *k* obtained in the SAT Mathematics test; and Z_{iks} refers to different indices of verbal ability of student *i* of ethnicity *k*. In the first version of model (1) the verbal ability indicator is the SAT Verbal score (*s*=1). The second version of model (1) uses the highest score between the R-SAT Verbal score and the original SAT score for minority students (*s*=2). Thus there are two versions of model (1) for African American students and two versions for Hispanic students for each academic year, which differed in the verbal ability index included (*s*=1 or *s*= 2). R-SAT was not available for White students therefore there was only one version of model (1) for each academic year, using just the SAT Verbal score, for them (*s*=1).[9] Finally e_i is a random error with expected value equal to 0 and variance equal to σ^2_e.

In model (2) $CUMUCGPA4_{iks}$ refers to the cumulative grade point average at the fourth college year of student *i* of ethnicity *k* considering verbal ability index *s*. In model (3) $GRAD4_{iks}$ is a binary variable indicating whether student *i* of ethnicity *k* graduated by the fourth year of college, considering verbal ability index *s*. For African American and Hispanic students, and just as in model (1), there were two versions of models (2) and (3), which differed in the verbal ability index included (*s*=1 or *s*= 2). For White students there was only one version of models (2) and (3), considering only the SAT Verbal score.

[9] The model presented in the text includes only SAT I Verbal (both the original and the maximum score between the SAT and the R-SAT scores) and SAT I Math scores as explanatory variables, and not SAT II scores, as (a) students took different SAT II tests, and the characteristics of these different tests vary considerably, and (b) most higher education institutions require only the SAT I exam and hence, results from these models will be more generalizable to other institutions. To check, we did conduct regressions including SAT II test scores as explanatory variables and found that they did not offer stronger evidence in support of the R-SAT Verbal test score, neither through larger and statistically significant coefficients nor through positive changes in the R^2. Details are available from the authors upon request.

All campus data are aggregated in the regression analyses and there is no control for the effect of discipline or campus on the dependent variable due to the small sample size of minority groups (Brown & Zwick, 2006). Student sample size also limited our ability to consider the within and between school variation in high-school GPA and API quintile (Zwick & Green, 2007), therefore no multilevel modeling was conducted.

The linear regressions analyses compared the explained variance across models measured by the standard R^2 (Singer & Willett, 2003). In logistic regression we used $R^2 = 1 - \{\frac{L(0)}{L(\hat{\theta})}\}^{2/n}$, where $L(0)$ is the likelihood of the intercept-only model, $L(\hat{\theta})$ is the likelihood of the specified model and *n* is the sample size. The standardized coefficients for the prediction of first-year GPA, 4th-year Cumulative GPA and 4th year graduation are shown in Appendix B.

Differential prediction of freshmen grades. Underprediction or overprediction is usually assessed by fitting one general prediction model for college students from all ethnic groups and then summing the regression residuals for a particular ethnic group. The average individual over or underprediction is calculated by adding the residuals and then dividing them by the number of students in each ethnic group. In this case, regression models 1.1 and 1.2 were estimated and the average residual by ethnic group compared. In this case the regression analyses did not distinguished among ethnicities. Two regressions were conducted and they differed only on the verbal ability indicator: the first one consider the SAT Verbal score (1.1) and the second one consider the maximum score between the SAT Verbal and the R-SAT Verbal score (1.2). All explanatory variables included in these models were described above.

$$UCGPA1_i = f_0 + f_1 APIQ_i + f_2 Educ_i + f_3 Inc_i + f_4 HSGPA_i + f_5 SATM_i + f_6 SATV_i + e_i \quad (1.1)$$

$$UCGPA1_i = l_0 + l_1 APIQ_i + l_2 Educ_i + l_3 Inc_i + l_4 HSGPA_i + l_5 SATM_i + l_6 Max(SATV\_RSATV)_i + e_i \quad (1.2)$$

Results

This section presents the results of this research in three parts: the calculation of the R-SAT, its predictive validity compared to the SAT and finally the R-SAT´s performance on the issue of overprediction and underprediction.

Freedle's Revised SAT Verbal Score

The adjusted scores were calculated for a total of 3,922 Hispanic examinees and 2,234 African American examinees who graduated from California public high schools. The R-SAT Verbal score mean is higher than the original SAT Verbal mean score in all of the ethnic groups and test forms that we examined. On average, the R-SAT Verbal score increases the mean score of African American students from 382.5 to 407 (6.4 percent increase) and the mean score of Hispanic students from 471.6 to 484.0 (2.6 percent increase). It is important to consider that the average SAT Verbal score for African American and Hispanic students is 439 and the Standard Deviation is 109 points (see Table 2). The increase between the mean SAT and R-SAT Verbal score, of 17 points, amounts to 16% of a standard deviation. Table 4 contrasts the original mean SAT Verbal score and the mean R-SAT Verbal score for Hispanic and African American examinees. These data are presented for the overall sample of Hispanic and African American examines, as well as for each test form in which the Freedle phenomenon could not be rejected.

Table 4
Mean Score for Minority Groups. Mean SAT Verbal Score Versus Mean R-SAT Verbal Score

Test Score	African American Students				Hispanic Students			
	Mean	n	Min	Max	Mean	n	Min	Max
				OVERALL SAMPLE				
SAT Verbal	382.52	2234	200.00	770.00	471.59	3922	200.00	800.00
R-SAT Verbal	406.97	2234	204.75	757.69	484.04	3922	319.17	782.73
				FORM IZ				
SAT Verbal	438.55	854	200.00	770.00	472.61	1904	200.00	790.00
R-SAT Verbal	468.58	854	334.00	757.69	486.43	1904	319.17	755.16
				FORM VD				
SAT Verbal	-	-	-	-	470.63	2018	200.00	800.00
R-SAT Verbal	-	-	-	-	481.79	2018	323.22	782.73
				FORM QI				
SAT Verbal	342.31	671	200.00	700.00	-	-	-	-
R-SAT Verbal	365.33	671	224.22	682.56	-	-	-	-
				FORM DX				
SAT Verbal	353.09	709	200.00	660.00	-	-	-	-
R-SAT Verbal	372.19	709	204.75	653.39	-	-	-	-

Table 5 provides greater detail about the degree to which the R-SAT Verbal score benefits minority students. Note that the bottom 3 rows display the students who benefit from the use of the R-SAT Verbal score. We observe that 68% of African American examinees (a total of 1,537 out of 2,234) improve their scores when the R-SAT Verbal score is considered in place of the SAT Verbal score. The same occurs for 58% of the Hispanic sample (a total of 2,271 over 3,922). In addition, the R-SAT Verbal tends to benefit mostly students in the low end of the original SAT Verbal score distribution. While most examinees increase their scores by between 0 and 50 points, the increment reaches as high as 202 points in a number of cases. On average, however, the score increase is not as large as Freedle described it to be (Freedle, 2003) and would be of little benefit to African Americans especially, who tend to start from a lower score, in comparison to Hispanics.

Table 5
Distribution of Score Difference by Ethnic Groups and Corresponding Mean SAT Verbal Score. Overall Sample

R-SAT Verbal Score Minus SAT Verbal Score (both end points included)	African American Examinees			Hispanic Examinees		
	Number	Percentage	Mean SAT Score	Number	Percentage	Mean SAT Score
[-106, -101]	-			2	0%	515.0
[-100, -51]	39	2%	433.6	95	2%	506.2
[-50, 0]	658	29%	438.7	1554	40%	518.4
[0, 49]	966	43%	396.2	1704	43%	468.9
[50, 101]	452	20%	301.6	418	11%	370.0
[100, 210]	119	5%	251.7	149	4%	276.1
TOTAL	2234	100%	382.5	3922	100%	471.6

The scatterplot of the SAT Verbal Score and R-SAT Verbal score (Figure 1) shows the same phenomenon. It is important to note the relative lower variance of R-SAT scores compared to SAT scores.

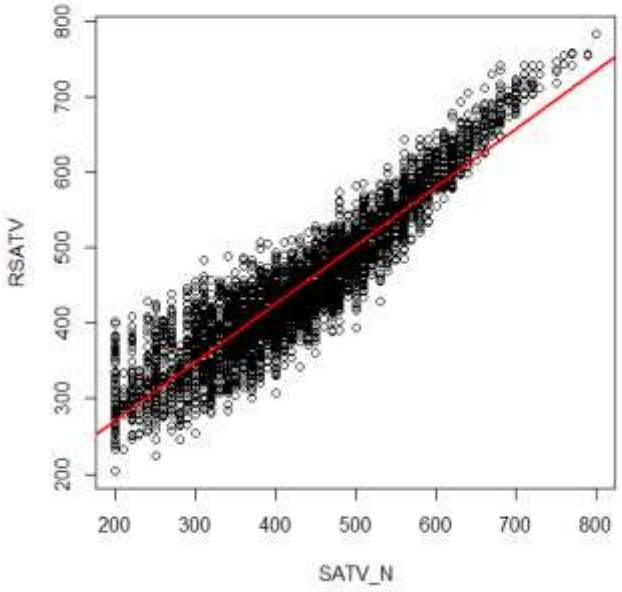

Figure 1. Scatterplot of R-SAT and SAT Verbal Score for all Hispanic and African American Students included in Sample.

In order to assess the impact of the revised SAT score in the admissions decisions of minority students, Freedle estimated and compared the number of African American students who would be offered admission at competitive colleges when considering each score. Freedle hypothesized that receiving an R-SAT score of at least 600 would be sufficiently meritorious to interest many colleges in an applicant who received such a score. Freedle chose to consider an SAT score of 600 or above as meritorious because students whose high school grade point average is between the 97th and 100th percentile receive an average SAT verbal score of 610 and, in addition, a

score of 600 also reflects a level of test performance that a small proportion of the test-taking population receives (Freedle, 2003). He reported that by considering the revised SAT score instead of the original SAT score, the number of African Americans scoring over 600 in two of the forms he analyzed increased from 166 to 235 (Form 4I) and from 117 to 167 (Form OB023) which he reported as equivalent to an increase in admission to selective colleges by 342% and by 334%, respectively (p. 15).

The analyses reported here show an effect in the same direction Freedle described. However, the impact in the number of African American students whose admissions are likely to have changed is smaller. When using the maximum of the SAT and the R-SAT Verbal scores, the number of African American students scoring over 600 increases from 79 to 86. This represents an increase of 8.9% over the original number of African American students in the sample scoring over 600 (see Table 6) or an increase from 3.5% of all African Americans to 3.8%. When considering both African American and Hispanic students, the number of students scoring over 600 increases from 458 (7.4% of all minority students) to 516 (8.3% of all minority students), which is equivalent to an increase of 12.6%. Overall, 7.4% of minority examinees now score over 600. In comparison, 3,889 White students, or 19.7% of all White examinees, score 600 or above and received an average score of 653.

Table 6
Number of Examinees Scoring 600, or Above, in the Sample and their Mean Scores

Ethnic Group	Number of Students Scoring Over 600 when considering SAT Verbal Score	Mean SAT Verbal	Number of Student Scoring Over 600 when considering Max. between SAT and R-SAT Verbal Score	Mean of Max. Between SAT V and R-SAT V	Total Number of Examinees in the Sample
African American Students	79 (3.5%)	637	86 (3.8%)	643	2,234
African American and Hispanic Students	458 (7.4%)	645	516 (8.3%)	648	6,156
White Students	3,889	653	-	-	19,778

The consideration of a different admission cut-off score other than 600, would only result in significant benefit for minorities if it was drastically reduced from 600. More than 60% of the African American and Hispanic students considered in this analysis would receive an R-SAT Verbal score below 450 therefore only an admission cut-off score around or below this level would result in a different admission decision. This sort of change in an admission score level, however, does not seem consistent with the assumption of being admitted to highly competitive colleges.

Predictive Validity of the Revised SAT Verbal Score

Table 7 shows the standard R^2 for the multivariate models estimated within each ethnic group.[10] The overall predictive power of the models examined varies depending on the academic outcome and ethnic group. In general, the models predict college grades better for White students than for minority students. While the capacity to predict annual college grades for White and Hispanic students tends to decline over time in college, the overall prediction of cumulative fourth-year grade point average is unexpectedly high for these same groups. In addition, the prediction of fourth year graduation is weaker than the prediction of annual college grades for both White and Hispanic students as well.

Table 7
Standard R^2 Multivariate Regression Model. 1999 Cohort. SAT Verbal Scores and the Maximum between the SAT Verbal Scores and the Revised SAT Verbal Scores

Model Includes:	African American Students	Hispanic Students	White Students	African American Students	Hispanic Students	White Students
Score	UCGPA 1st Year			UCGPA 2nd Year		
SAT V	9.8%	16.2%	21.5%	8.5%	14.1%	16.8%
Max [SATV or R-SATV]	9.3%	15.9%	-	8.4%	13.4%	-
N	78	597	2253	73	540	2120
	UCGPA 3rd Year			UCGPA 4th Year		
SAT V	5.1%	9.2%	13.2%	13.9%	6.2%	13.4%
Max [SATV or R-SATV]	4.4%	8.4%	-	14.0%	5.6%	-
N	67	497	1964	64	476	1904
	UC CUM GPA 4th YEAR			UC GRADUATION BY 4th		
SAT V	9.5%	16.2%	20.9%	9.5%	9.4%	5.2%
Max [SATV or R-SATV]	10.0%	15.4%	-	9.0%	9.2%	-
N	65	481	1927	78	613	2314

Note: Pseudo R^2 is reported for the logistic regression used to predict fourth-year graduation.

[10] In order to increase the sample size, results for the R-SAT Verbal score were combined across all SAT forms. This aggregation was possible because the performance in each form was previously scaled by ETS. Scaling refers to a psychometric process conducted to achieve comparability among test score from different test forms. The aggregation conducted also assumes that the four SAT forms were equated during test development. For an introduction to traditional scaling and equating methods see Kolen (1988). Equating is a process different from scaling and aims to adjust for differences in difficulty among test forms.

Table 7 shows that the capacity to predict college outcomes using the R-SAT Verbal score is close to, but slightly less, than the predictive power achieved when using the original SAT score. The model using the R-SAT Verbal score predicts better than the model using the original SAT score only in two out of twelve cases - just for the African American group´s fourth year college grade point average and fourth year cumulative grade point average. The small differences in predictive power are also not of large practical significance as they ranged between 0.1 and 0.5 percentage points.

Differential Prediction of Freshmen Grades

We find underprediction of White students' grades (0.01) and overprediction of Hispanic (-0.025) and African American students' first year grades (-0.098) when using the SAT, just as previous research did (Ramist et al., 1994; Ramist, Lewis, & McCamley-Jenkins, 2001). We found no improvement in the prediction power from using the R-SAT Verbal score for minority groups. On the contrary, the prediction errors for minorities decreased (increased in absolute terms) when using the maximum from the SAT and R-SAT Verbal score, to (-0.032) for Hispanic students and to (-0.114) for African American students respectively. The same analysis was conducted for fourth-year cumulative UC GPA and the average underprediction for African American and Hispanic students increased in absolute terms as well (from -0.181 to -0.194 and from -0.033 to -0.040 respectively).[11] These differences are larger and thus more important than differences in the prediction of first- year GPA.

Discussion and Conclusion

Analyses presented above show that in the sample the R-SAT score does result in increased scores for minority students, although not as much as Freedle expected. On average, it increases scores by 24 points (6%) for African American students and by 12 points (2.5%) for Hispanic students. Using Freedle's assumptions, the consideration of the R-SAT would change admissions decisions for minority students admitted into selective colleges by about 10%. This is much less than Freedle´s prediction of approximately 300% increase but it should be given some consideration since such an increase could be educationally significant in some contexts, especially at the most selective institutions. The small increases in R-SAT scores reported in our research are consistent with the magnitude of score increase reported by Dorans (2004) and Dorans and Zeller (2004a). In addition, the predictive validity analyses show virtually no difference in the capacity to predict short and long-term outcomes when using either the original or the revised SAT score. The R^2 using the SAT Verbal score for the prediction of college grades for African American, Hispanic and White students are consistent with the results reported by similar studies (Geiser & Studley, 2002; Zwick et al., 2004). Geiser and Studley (2002), for example, reported R^2s close to 10% for African American students (pp. 15). When predicting graduation, however, the models predict better for African Americans and Hispanics than for White students. The limited incremental predictive power of the maximum score between the R-SAT Verbal and the Verbal scores may be explained by the lower variance observed in R-SAT scores when compared to SAT Verbal scores, which is related to the fact that R-SAT scores are actually regression predictions.

Also, results show that the traditional problem of overprediction and underprediction would remain approximately the same when using the revised SAT score. On average, the overprediction estimated in this study lays in the range between the overprediction reported for African American

[11] Regression results are available from authors upon request.

students by Geiser and Studley (2002) and by Zwick, Brown, and Sklar (2004)[12] and the overprediction Ramist et al. (1994) reported for the same group of students (-0.16). For Hispanic students the overprediction is smaller than the one reported by Ramist et al. (2001) (-0.13) and similar to some of the results reported by Zwick et al. (2004; see for example Berkeley 1996–1997 mega-cohort, Irvine 1998–1999 mega-cohort, San Diego 1996–1997 mega-cohort).

This research has several limitations. Among them is the fact that predictive validity analyses were conducted on a group of students who were already accepted to college and therefore present a significant restriction of range in some of the explanatory variables. In addition, many students who did not attend selective colleges might have matriculated at such schools if their R-SAT scores had been used in the admission process, which limits in some extent the validity of our findings. This limitation, however, has also been the case in other predictive validity studies (Geiser & Studley, 2002; Zwick, 2002; Zwick et al., 2004; Zwick & Sklar, 2005). The aggregation of different ethnic groups in order to obtain the R-SAT scores is still subject to Dorans and Zeller´s original criticisms (Dorans & Zeller, 2004a, 2004b). Recent changes to the content of the SAT and the inclusion of a Writing test may limit the generalizability of the findings presented here since they were based in somewhat older test forms. Larger sample size for each minority group may be desirable in order to implement future research, especially for African American students. Increasing the sample size, however, remains a daunting task as it requires data from an even greater number of colleges and universities than the nine campuses of the University of California examined here. Furthermore, and despite the limited sample size of African American and Hispanic students, we were still able to observe results that were similar to those reported by previous research, such as the statistical significance and practical importance of high school grades for predicting college grades and graduation (see Appendix B). These results provide support for the validity of our results for these particular samples.

We think it is important to highlight the consistency of the results obtained in the numerous and diverse analyses implemented in this research: no strong evidence in favor of the R-SAT score is observed when (a) recalculating the scores using only the most difficult items for minorities, (b) when using that maximum between the R-SAT and the SAT Verbal score to directly predict short and long term outcomes for African American and Hispanic students using models that did not considered SAT II scores, and (c) when evaluating the overprediction and underprediction problem for African American and Hispanic students. Although not presented here, we also found results that did not support the use of the R-SAT score nor the maximum between the R-SAT and the SAT Verbal score in models that considered SAT II scores and when using models that did not control for school quality and allowed us to have larger sample sizes.

The findings presented in this article consistently reveal that there are only quite minimal benefits associated with Freedle's R-SAT and suggest that, rather than using measures aimed to complement the SAT, efforts and energy should be directed to studying the phenomenon behind the systematic relationship between item difficulty and DIF estimates and directly addressing those issues during test development. The investigation of potential causes should explore Freedle's proposed explanation, the influence of academic versus home language (Freedle, 2010) —including examination of the cognitive processes of students while taking the test as well as quantitative analyses and modeling techniques (De Boeck, 2010). In addition, further research should investigate the relationship between Freedle's phenomenon and alternative forms of guessing such as differential guessing strategies between White students and students from other ethnic groups.

[12] Except for the 1998–1999 UCLA "mega-cohort" for the African American group. We focused our attention on Zwick et al.´s model 6, which is the most similar to the analyses reported in this section.

These results also suggest that alternative policy options should be considered if the goal is to increase the representation of minority groups in higher education, especially at highly selective institutions (Bowen et al., 2009). Those options may include the use of school quality indices as input in the admissions processes (Zwick & Himelfarb, 2011) as well as explicitly considering nonacademic outcomes as college goals and therefore adjusting the weight of admission indicators accordingly (Sinha et al., 2011).

References

Allen, J., Robbins, S., & Sawyer, R. (2010). Can measuring psychosocial factors promote college success? *Applied Measurement in Education*, *23*(1), 1-22. http://dx.doi.org/10.1080/08957340903423503

Astin, A. W., Tsui, L., & Avalos, J. (1996). *Degree attainment rate at American colleges and universities: effect of race, gender and institutional type*. Los Angeles, CA: Higher Education Research Institute, Graduate School of Education & Information Studies.

Bean, J. P., & Mentzer, B. S. (1985). A conceptual model of nontraditional undergraduate student attrition. *Review of Educational Research, 55*(4), 485-540. http://dx.doi.org/10.3102/00346543055004485

Bowen, W., & Bok, D. (1998). *The shape of the river. Long-term consequences of considering race in college and university admissions*. Princeton, NJ: Princeton University Press.

Bowen, W., Chingos, M., & McPherson, M. (2009). *Crossing the Finish line. Completing college at America´s public universities*. Princeton, NJ: Princeton University Press.

Breland, H., Maxey, J., Gernand, R., Cumming, T., & Trapani, C. (2001). *Trends in college admission 2000: A report of a national survey of undergraduate admissions policies, practices, and procedures*. Retrieved from http://www.semworks.net/aboutus/resources/docs/trends_in_college_admission.pdf

Bridgeman, B., & Burton, N. (2005). *Does scoring only the hard questions on the SAT make it fairer?* Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada.

Bridgeman, B., Pollack, J., & Burton, N. (2004). *Understanding what SAT Reasoning Test Scores add to high school grades: A straightforward approach* (No. 2004-04). New York, NY: College Board.

Brown, T., & Zwick, R. (2006). *Using Hierarchical Linear Models to describe first-year grades at the University of California*. Paper presented at Annual Meeting of the American Educational Research Association, San Francisco, CA.

Burton, N., & Ramist, L. (2001). *Predicting success in college: SAT studies of classes graduating since 1980* (No. 2001-2). New York, NY: College Board.

California Department of Education (2014). *Status of the Academic Performance Index and the 3-year average. Information guide*. Retrieved from http://www.cde.ca.gov/ta/ac/ap/documents/infoguide14.pdf

Camara, W., & Echternacht, J. (2000). *The SAT I and high school grades: Utility in predicting success in college*. (No. Research Report RN-10). New York, NY: College Entrance Examination Board.

Camara, W., & Kimmel, E. W. (Eds.). (2005). *Choosing students: Higher education admissions tools for the 21st Century*. Mahwah, NJ: Lawrence Erlbaum.

Camara, W., & Sathy, V. (2004). *College Board response to Harvard educational review article by Freedle*. Retrieved from http://www.collegeboard.com/research/pdf/051425Harvard_050406.pdf#search=%22college%20board%20freedle%22

Camilli, G. (2006). Test fairness. In R. Brennan (Ed.) *Educational measurement* (4th ed., p. 221-246). Washington, DC: American Council on Education/Praeger.

Cleary, T. A. (1968). Test bias: Prediction of Negro and White students in integrated colleges. *Journal of Educational Measurement, 5*(2), 115-124. http://dx.doi.org/10.1111/j.1745-3984.1968.tb00613.x

De Boeck, P. (2010). *Another look at bias in the SAT* [Web log post]. Retrieved from http://www.hepg.org/blog/45

Diaz-Guerrero, R., & Szalay, L. B. (1991). *Understanding Mexicans and Americans: Cultural perspectives in conflict*. New York, NY: Plenum Press. http://dx.doi.org/10.1007/978-1-4899-0733-2

Dorans, N. (2004). Freedle's table 2: Fact or fiction. *Harvard Educational Review*, *74*(1), 62-72.

Dorans, N. (2010). *Unfair treatment vs. Confirmation bias? Comments on Santelices & Wilson. ETS-RR-10-20*. New Jersey: ETS. http://dx.doi.org/10.17763/haer.74.1.8729105044552127

Dorans, N., & Holland, P. (1992). *DIF detection and description: Mantel-Haenszel and standardization* (No. RR-92-10). Princeton, NJ: Educational Testing Services.

Dorans, N., & Kulick, E. (1983). *Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December 1977: An application of the standardization approach* (No. RR-83-9). Princeton, NJ: Educational Testing Service.

Dorans, N., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, *23*(4), 355-368. http://dx.doi.org/10.1111/j.1745-3984.1986.tb00255.x

Dorans, N., & Zeller, K. (2004a). *Examining Freedle's claims and his proposed solution: Dated data, inappropriate measurement, and incorrect and unfair scoring* (No. RR-04-26). Princeton, NJ: Educational Testing Service.

Dorans, N., & Zeller, K. (2004b). *Using score equity assessment to evaluate the equitability of the hardest half of a test to the total test* (No. RR-04-43). New Jersey, NJ: Educational Testing Service.

FairTest (2003). *SAT I: A faulty instrument for predicting college success*. Retrieved from http://fairtest.org/facts/satvalidity.html

Frary, R. B. (1988). Formula scoring of multiple-choice tests (Correction for guessing). *Educational Measurement: Issues and Practice*, *7*(2), 33-38. http://dx.doi.org/10.1111/j.1745-3992.1988.tb00434.x

Freedle, R. (2003). Correcting the SAT's ethnic and social-class bias: A method for reestimating SAT scores. *Harvard Educational Review*, *73* (1), 1-43. http://dx.doi.org/10.17763/haer.73.1.8465k88616hn4757

Freedle, R. (2010). On replicating ethnic test bias effects: The Santelices and Wilson study. *Harvard Educational Review*, *80*, 394-404. http://dx.doi.org/10.17763/haer.80.3.l050025058204016

Geiser, S., & Santelices, M. V. (2007). Validity of high school grades in predicting student success beyond the freshman year: High-school record vs. standardized tests as indicators of four-year college outcomes. In *Center for Studies on Higher Education, University of California, Berkeley*. Research & Occasional Paper Series, CSHE.6.07, 1-35. Retrieved from http://cshe.berkeley.edu/publications/docs/ROPS.GEISER._SAT_6.12.07.pdf

Geiser, S., & Studley, R. (2002). UC and the SAT: Predictive validity and differential impact of the SAT I and SAT II at the University of California. *Educational Assessment*, *8*(1), 1-26. http://dx.doi.org/10.1207/S15326977EA0801_01

Hezlett, S. A., Kuncel, N. R., Vey M., Ahart, A. M., Ones, D. S., Campbell, J. P., & Camara, W. J. (2001). *The effectiveness of the SAT in predicting success early and late in college: A comprehensive meta-*

analysis. Paper presented at the Annual Meeting of the National Council of Measurement in Education, Seattle, USA.

Holland, P., & Wainer, H. (1993). Concluding remarks and Suggestions. In P. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp. 419-422). Hillside, NJ: Lawrence.

Hout, M. (2012). Social and economic returns to college education in the United States. *Annual Review of Sociology, 38*(1), 379-400. http://dx.doi.org/10.1146/annurev.soc.012809.102503

Kobrin, J. L., Patterson, B. F., Shaw, E. J., Mattern, K. D., & Barbuti, S. M. (2008). *Validity of the SAT for predicting first-year college grade point average* (College Board Research Rep. No. 2008-5). New York, NY: The College Board.

Kolen, M. (1988). Traditional equating methodology. *Educational Measurement: Issues and Practice, 7(*4*),* 29-36. http://dx.doi.org/10.1111/j.1745-3992.1988.tb00843.x

Kulick, E., & Hu, P. G. (1989). *Examining the relationship between differential item functioning and item difficulty* (College Board Report No. 89-5; ETS No. RR-89-18). New York: College Entrance Examination Board.

Kyllonen, P. C. (2008). *The research behind the ETS Personal Potential Index*. Retrieved from the ETS website http://www.ets.org/Media/Products/PPI/10411_PPI_bkgrd_report_RD4.pdf

Mattern, K., & Patterson, B. (2009). *Is performance on the SAT related to college retention?* (College Board Research Report No. 2009-7). New York: College Board.

Mattern, K., & Patterson, B. (2011a). *The relationship between SAT scores and retention to the third year: 2006 SAT validity sample* (College Board Statistical Report No. 2011-2). New York, NY: College Board.

Mattern, K., & Patterson, B. (2011b). *The relationship between SAT scores and retention to the fourth year: 2006 SAT validity sample* (College Board Statistical Report No. 2011-6). New York, NY: College Board.

Oswald, F., Schmitt, N., Kim, B. H., Ramsay, L., & Gillespie, M. (2004). Developing a biodata measure and situational judgment inventory as predictors of college student performance. *Journal of Applied Psychology*, *89*(2), 187-207. http://dx.doi.org/10.1037/0021-9010.89.2.187

Pascarella, E. T., & Terenzini, E. T. (1991). *How college affects students: Findings and in-sights from twenty years of research.* San Francisco: Jossey-Bass.

Perfetto, G. (1999). *Toward taxonomy of the admissions decision-making process: A public document based on the first and second College Board conferences on admissions models.* New York, NY: College Board.

Ramist, L., Lewis, C., & McCamley-Jenkins, L. (1994). *Student group differences in predicting college grades: Sex, language and ethnic groups* (No. Report No. 93-1). New York, NY: College Entrance Examination Board.

Ramist, L., Lewis, C., & McCamley-Jenkins, L. (2001). *Using achievement Tests/SAT II: Subject tests to demonstrate achievement and predict college grades: Sex, language, ethnic, and parental education groups* (No. 2001-05). New York, NY: College Board.

Ramsey, P. (1993). Sensitivity review: The ETS experience as a case study. In P. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp. 367-388). Hillsdale, NJ: Lawrence Erlbaum Associates.

Reason, R. (2009). An examination of the persistence research through the lens of a comprehensive conceptual framework. *Journal of College Student Development*, *50*(6), 659-682. http://dx.doi.org/10.1353/csd.0.0098

Robbins, S. B., Lauver, K., Le, H., Davis, D., & Langley, R. (2004). Do psychological and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin, 130*(2), 261-288. http://dx.doi.org/10.1037/0033-2909.130.2.261

Rogers, H. (1999). Guessing in multiple choice tests. In G. Masters, & J. Keeves (Eds.), *Advances in measurement in educational research and assessment* (pp. 235-243). Amsterdam: Pergamon.

Rothstein, J. M. (2004). College performance predictions and the SAT. *Journal of Econometrics*, *121*(1-2), 297-317. http://dx.doi.org/10.1016/j.jeconom.2003.10.003

Santelices, M. V., & Wilson, M. (2010a). Responding to claims of misrepresentation. *Harvard Educational Review, 80*(3), 413-416. http://dx.doi.org/10.17763/haer.80.3.p35268145347xp56

Santelices, M. V., & Wilson, M. (2010b). Unfair treatment? The case of Freedle, the SAT, and the standardization approach to differential item functioning. *Harvard Educational Review, 80*(1), 106-134. http://dx.doi.org/10.17763/haer.80.1.j94675w001329270

Santelices, M. V., & Wilson, M. (2012). On the relationship between differential item functioning and item difficulty: An issue of methods? Item response theory approach to differential item functioning. *Educational and Psychological Measurement,* 72(1), 5-36. http://dx.doi.org/10.1177/0013164411412943

Schmitt, A., & Bleistein, C. (1987). *Factors affecting differential item functioning for black examinees on Scholastic Aptitude Test analogy items* (No. RR-87-23). Princeton, NJ: Educational Testing Service.

Singer, J., & Willett, J. (2003). *Applied longitudinal data analysis.* New York: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780195152968.001.0001

Sinha, R., Oswald, F., Imus, A., & Schmitt, N. (2011). Criterion-focused approach to reducing adverse impact in college admissions. *Applied Measurement in Education*, *24*(2), 1-25. http://dx.doi.org/10.1080/08957347.2011.554605

Steele, C. M., & Aronson, J. (1998). How stereotypes influence the standardized test performance of talented African American students. In C. Jencks, & M. Phillips (Eds.), *Black-White test score differences* (pp. 401-427). Washington, DC: Brookings.

Stemler, S. E. (2012). What should university admissions tests predict? *Educational Psychologist, 47*(1), 5-17. http://dx.doi.org/10.1080/00461520.2011.611444

Sternberg, R. J. (1999). A triarchic approach to the understanding and assessment of intelligence in multicultural populations. *Journal of School Psychology, 37*(2), 145–159. http://dx.doi.org/10.1016/S0022-4405(98)00029-6

Sternberg, R. J. (2003). Our research program validating the triarchic theory of successful intelligence: Reply to Gotfredson. *Intelligence, 31*, 399-413. http://dx.doi.org/10.1016/S0160-2896(02)00143-5

Sternberg, R. J., Gabora, L., & Bonney, C. R. (2012). Introduction to the special issue on college and university admissions. *Educational Psychologist, 47*(1), 1-4. http://dx.doi.org/10.1080/00461520.2011.639652

Tinto, V. (2006). Research and practice of student retention: What next? *Journal of College Student Retention, 8*(1), 1-19. http://dx.doi.org/10.2190/4YNU-4TMB-22DJ-AN4W

Wainer, H. (2009). Ethnic bias or statistical artifact? Freedle´s folly. In H. Wainer (Ed.), *Picturing the uncertain world. How to understand, communicate and control uncertainty through graphical display*. New Jersey: Princeton University Press.

Willingham, W. (1985). *Success in college: The role of personal qualities and academic ability*. New York, NY: College Entrance Examination Board.

Willingham, W., Lewis, C., Morgan, R., & Ramist, L. (1990). *Predicting college grades: An analysis of institutional trends over two decades*. New York, NY: College Entrance Examination Board and Educational Testing Service.

Wilson, K. (1983). *A review of research on the prediction of academic performance after the freshman year* (No. 83-2). New York, NY: College Board.

Young, J. (2004). Differential validity and prediction: Race and sex differences in college admissions testing. In R. Zwick (Ed.), *Rethinking the SAT. The future of standardized testing in university admissions* (pp. 289-301). New York, NY: Routledge Falmer.

Zwick, R. (2002). *Fair game? The use of standardized admissions tests in higher education*. New York: Routledge Falmer.

Zwick, R. (2006). Higher education admissions testing. In R. L. Brennan (Ed.) *Educational measurement* (4th ed., p. 221-246). Westport, CT: American Council on Education/Praeger.

Zwick, R., Brown, T., & Sklar, J. C. (2004). *California and the SAT: A reanalysis of University of California admissions data.* Research & Occasional Paper Series: CSHE.8.04. Berkeley, CA: University of California, Berkeley. Retrieved from http://cshe.berkeley.edu/publications/publications.php?s=1

Zwick, R., & Green, J. (2007). New perspectives on the correlation of SAT scores, high school grades and socioeconomic factors. *Journal of Educational Measurement, 44*(3), 23-45. http://dx.doi.org/10.1111/j.1745-3984.2007.00025.x

Zwick, R., & Himelfarb, I. (2011). The effect of high school socioeconomic status on the predictive validity of SAT scores and high school grade-point average. *Journal of Educational Measurement*, *48*(2), 101-121. http://dx.doi.org/10.1111/j.1745-3984.2011.00136.x

Zwick, R., & Schlemer, L. (2004). SAT validity for linguistic minorities at the University of California, Santa Barbara. *Educational Measurement: Issues and Practice*, *23*(1), 6-16. http://dx.doi.org/10.1111/j.1745-3992.2004.tb00148.x

Zwick, R., & Sklar, J. (2005). Predicting college grades and degree completion using high school grades and sat scores: The role of student ethnicity and first language. *American Educational Research Journal, 42*(3), 439-464. http://dx.doi.org/10.3102/00028312042003439

Appendix A
Predictive Validity of R-SAT. Alternative Models

This appendix presents the results of alternative models to evaluate the relative predictive capacity of the revised SAT Verbal score. These alternative models consider the same explanatory variables as presented in the text except for school ranking (API scores). In addition, these models compare the predictive capacity of the original SAT Verbal score to that of the R-SAT score and not only to that of the maximum score between the original SAT and the R-SAT score.[13] The R-SAT was only available for minority students. The indicator of secondary education quality is excluded from the analyses in order to increase the sample size, especially that of African American students, and in order to avoid the potential clustering of errors at the school level. Just as in the main text, the capacity to predict short and long term academic outcomes is analyzed by ethnic group and academic outcome. In order to conduct the analyses, all SAT forms and ethnic groups in which there was evidence of DIF were combined, therefore the analyses includes Hispanic students who took the 1999 forms (IZ and VD) and African American taking the forms 1999 (IZ) and 1994 forms (QI and DX).

These models result in findings similar to the ones reported in the main text and do not provide stronger evidence in favor of using either the R-SAT nor the Max SATV_R-SATV for minority students. There are only three cases out of twelve in which one of the alternative measures predicts better than the SAT Verbal score. Two of those cases refer to the prediction of college grades for African American students (first-year and fourth-year GPA) and the difference in overall predictive power ranges between 0.20% and 0.30% (see Table 1). The third case occurs when predicting fourth-year graduation for Hispanic students and only in this case the difference in explained variance when using the original and R-SAT Verbal Score (1.13%) seems to have some practical meaning.[14] However when analyzing the standardized coefficients, only in the prediction of fourth year grades for Hispanic students there is an increase in the coefficient associated to the SAT V score and this is statistically significant. In all other regressions, the coefficient associated to the alternative versions of the SAT Verbal score are either not statistically significant (first-year and fourth-year GPA for African American students) or smaller than the coefficient associated to the original SAT Verbal score (fourth-year graduation for Hispanic students).

In summary, and although there are some cases in which we observe weak evidence in favor of the R-SAT Verbal Scores and the Max SATV_R-SATV, none of the alternative models considered in this appendix provide strong support for using these modified admission scores.

[13] In this case, the R-SAT Verbal Score can be positive or negative.

[14] Still, even if the R-SAT Verbal Score probed to be more predictive than the original SAT V scores it is unclear what would be the face validity of using the R-SAT Verbal Score, a variable that can benefit or harm minority students who are already affected by DIF in the SAT Verbal items.

Table

Adjusted R^2 of the Original and Revised SAT Verbal scores. Alternative Multivariate Regression Models

Score	African American Students	Hispanic Students	White Students	African American Students	Hispanic Students	White Students
	UCGPA 1st Year			UCGPA 2nd Year		
SAT V	8.00%	14.52%	20.43%	5.12%	13.15%	15.78%
Max [SATV or R-SATV]	8.28%	14.15%	-	4.87%	12.36%	-
R-SAT Verbal Score	8.08%	13.95%	-	4.92%	11.48%	-
N	191	603	2879	175	545	2688
Score	UCGPA 3rd Year			UCGPA 4th Year		
SAT V	6.99%	7.86%	7.07%	7.52%	5.05%	11.66%
Max [SATV or R-SATV]	6.92%	7.05%	-	7.69%	4.41%	-
R-SAT Verbal Score	6.91%	7.02%	-	7.30%	3.97%	-
N	156	503	2479	149	482	2416
Score	UC CUM GPA 4th YEAR			UC GRADUATION BY 4th YEAR*		
SAT V	10.01%	15.32%	19.79%	10.29%	11.90%	7.23%
Max [SATV or R-SATV]	9.93%	14.46%	-	9.83%	11.67%	-
R-SAT Verbal Score	10.00%	13.58%	-	9.91%	11.66%	-
N	155	487	2449	199	619	3009

Note: Pseudo R^2 is reported for the logistic regression used to predict fourth-year graduation.

Appendix B

Prediction Results of SAT and Maximum between SAT Verbal Score and R-SAT Score. Standardized Coefficients

Table 1
Prediction of First-Year UC GPA. Standardized Estimates and Statistical Significance (p-values) by Ethnic Group.

Regression Model	API Quintile	Parents Education	Income Level	HS GPA	SAT Math	SAT Verbal	Max [SATV or R-SAT Verbal Score]	Standard R^2	N
				HISPANIC STUDENTS					
1.1	0.093	0.057	0.082	0.306	-0.065	0.152		16.2%	597
	(0.035)	*(0.223)*	*(0.056)*	*(<.0001)*	*(0.180)*	*0.001*			
1.2	0.095	0.059	0.082	0.308	-0.054		0.131	15.9%	597
	(0.031)	*(0.215)*	*(0.056)*	*(<.0001)*	*(0.266)*		*(0.005)*		
				AFRICAN AMERICAN STUDENTS					
1.1	0.087	0.118	-0.103	0.069	0.067	0.156		9.8%	78
	(0.513)	*(0.369)*	*(0.431)*	*(0.572)*	*(0.678)*	*(0.325)*			
1.2	0.092	0.120	-0.102	0.071	0.091		0.122	9.3%	78
	(0.495)	*(0.367)*	*(0.434)*	*(0.567)*	*(0.565)*		*(0.433)*		
				WHITE STUDENTS					
1.1	0.087	0.082	0.031	0.346	-0.015	0.192		21.5%	2253
	(<.0001)	*(<.0001)*	*(0.115)*	*(<.0001)*	*(0.516)*	*(<.0001)*			

Table 2

Prediction Power of Cumulative Fourth-Year UC GPA. Standardized Estimates and Statistical Significance (p-values) by Ethnic Group

Regression Model	API Quintile	Parents Education	Income Level	HS GPA	SAT Math	SAT Verbal	Max [SATV or R-SAT Verbal Score]	Standard R^2	N
				HISPANIC STUDENTS					
2.1	0.006	0.024	0.114	0.221	-0.096	0.314		16.2%	481
	(0.910)	*(0.653)*	*(0.017)*	*(<.0001)*	*(0.075)*	*(<.0001)*			
2.2	0.007	0.025	0.113	0.221	-0.078		0.288	15.4%	481
	(0.883)	*(0.634)*	*(0.018)*	*(<.0001)*	*(0.145)*		*(<.0001)*		
				AFRICAN AMERICAN STUDENTS					
2.1	0.221	-0.004	-0.048	0.224	-0.158	0.113		9.5%	65
	(0.120)	*(0.978)*	*(0.736)*	*(0.1041)*	*(0.355)*	*(0.507)*			
2.2	0.232	-0.019	-0.045	0.213	-0.179		0.153	10.0%	65
	(0.106)	*(0.893)*	*(0.753)*	*(0.124)*	*(0.295)*		*(0.377)*		
				WHITE STUDENTS					
2.1	0.078	0.066	0.014	0.352	-0.081	0.241		20.9%	1927
	(0.000)	*(0.002)*	*(0.513)*	*(<.0001)*	*(0.001)*	*(<.0001)*			

Table 3
Prediction Power of Fourth-Year UC Graduation. Standardized Estimates and Statistical Significance (p-values) by Ethnic Group

Regression Model	API Quintile	Parents Education	Income Level	HS GPA	SAT Math	SAT Verbal	Max [SATV or R-SAT Verbal Score]	Standard R^2	N
				HSPANIC STUDENTS					
3.1	0.140	-0.021	0.295	0.234	-0.111	0.134		9.4%	613
	(0.019)	*(0.754)*	*(0.000)*	*(<.0001)*	*(0.094)*	*(0.038)*			
3.2	0.142	-0.020	0.296	0.235	-0.103		0.115	9.2%	613
	(0.018)	*(0.767)*	*(0.000)*	*(<.0001)*	*(0.122)*		*(0.069)*		
				AFRICAN AMERICAN STUDENTS					
3.1	-0.266	0.390	-0.005	0.168	-0.355	0.297		9.5%	78
	(0.249)	*(0.079)*	*(0.982)*	*(0.378)*	*(0.178)*	*(0.224)*			
3.2	-0.253	0.390	-0.003	0.163	-0.310		0.241	9.0%	78
	(0.269)	*(0.079)*	*(0.988)*	*(0.397)*	*(0.220)*		*(0.303)*		
				WHITE STUDENTS					
3.1	0.081	0.031	0.057	0.233	-0.069	0.078		5.2%	2314
	(0.001)	*(0.220)*	*(0.028)*	*(<.0001)*	*(0.016)*	*(0.005)*			

Note: Pseudo R^2 is reported for the logistic regression used to predict fourth-year graduation.

About the Authors

Maria Veronica Santelices
Facultad de Educación, Pontificia Universidad Católica de Chile
vsanteli@uc.cl
Maria Veronica Santelices, PhD, is an associate professor at Pontificia Universidad Católica de Chile, Department of Education. Her research interests include educational measurement and educational policy. She received a PhD in Education from University of California Berkeley in 2007.
http://orcid.org/0000-0002-9659-8670

Mark Wilson
University of California, Berkeley
markw@berkeley.edu
Mark Wilson, PhD, is a professor of education at the University of California, Berkeley, where he teaches courses on measurement in the social sciences, multidimensional measurement and applied statistics. He was the president of the Psychometric Society for 2011—12, and also became a member of the U.S. National Academy of Education in the same year. He has chaired two US National Research Council committees.
http://orcid.org/0000-0002-0425-5305

education policy analysis archives

Volume 23 Number 113 November 16th, 2015 ISSN 1068-2341