Educational and Psychological Measurement
72(1) 5–36

DOI: 10.1177/0013164411412943
http://epm.sagepub.com

On the Relationship Between Differential Item Functioning and Item Difficulty: An Issue of Methods? Item Response Theory Approach to Differential Item Functioning

María Verónica Santelices[1] and Mark Wilson[2]

Abstract

The relationship between differential item functioning (DIF) and item difficulty on the SAT is such that more difficult items tended to exhibit DIF in favor of the focal group (usually minority groups). These results were reported by Kulick and Hu, and Freedle and have been enthusiastically discussed by more recent literature. Examining the validity of the original reports of this systematic relationship is important so that we can move on to investigating more effectively its causes and the consequences associated to test score use. This article explores the hypothesis that the observed relationship between DIF and item difficulty observed in the SAT could be because of one of the following explanations: (a) the confounding of DIF and impact by the shortcomings of the standardization approach and/or (b) by random guessing. The relationship between DIF and item difficulty is examined using item response theory, which better controls for differences between impact and DIF than the standardization approach and also allows us to test the importance of guessing. The results obtained generally find evidence in support of the relationship between item difficulty and DIF suggesting that the phenomenon reported by earlier research is not a mere artifact of the statistical methodologies used to study DIF.

Keywords

differential item functioning, bias, SAT, item response theory, standardization approach

[1]Pontificia Universidad Católica de Chile, Santiago, Chile
[2]University of California Berkeley, Berkeley, USA

Corresponding Author:
María Verónici Santelices, Pontificia Universidad Católica de Chile, Hernando de Aguirre 959 Depto 307, Providencia, Santiago, Chile
Email: vsanteli@uc.cl

Introduction

The objective of this article is to explore reports of unfair results for African American and other minority students on the SAT. Specifically, we studied the correlation between item difficulty and differential item functioning (DIF) values (i.e., African Americans tend to perform better in harder items and White students perform better in easier items) reported by Kulick and Hu (1989) and Freedle (2003). Both investigations used the standardization approach to study DIF.[1] In contrast, we will use item response theory (IRT) methods.

Previously, researchers have offered at least two types of accounts to explain this empirical finding: (a) the contention that the standardization approach does not sufficiently control for differences in the mean ability of the two compared distributions (Dorans, 2004; Schmitt & Bleistein, 1987), and (b) the guessing behavior observed in multiple-choice items (Kulick & Hu, 1989; Schmitt & Bleistein, 1987; Wainer, 2009; Wainer & Skorupski, 2005).

IRT methods, in general, allow one to better control for differences in the mean ability level of examinees because, when the IRT model fits, item parameters are less confounded with sample characteristics (Hambleton, Swaminathan, & Rogers, 1991) than in observed-score DIF methods. Furthermore, the explicitness and flexibility of the IRT models are convenient for exploring different hypotheses regarding the relationship between item difficulty and DIF estimates such as the role of guessing.

The one-parameter logistic (1PL) model, one model used in this article, is conceptually similar to the standardization approach and allows us to test whether the relationship between item difficulty and DIF estimates is affected by the methodology used. In addition, we also used a second model, the 3PL model, which is a more complex IRT model that explicitly incorporates the possibility that low-ability students respond to an item correctly by chance. Exploring DIF using this latter model will shed light on the role of guessing in the relationship between item difficulty and DIF estimates.

Scherbaum and Goldstein (2008) found evidence supporting the generalizability of the correlation between item difficulty and DIF described by Freedle (2003) and Kulick and Hu (1989) using IRT methods. However, they used the 2PL model, which does not incorporate the possibility of guessing, and analyzed a different standardized test than the one analyzed by Freedle (they used a test of ninth graders' knowledge of U.S. civics).

Our work will further contribute to explore the role of methods in explaining the relationship between item difficulty and DIF using IRT models and data from SAT administrations.

Background

The relationship between item difficulty and DIF statistics has been one of the foci of SAT research since the 1970s (Carlton & Harris, 1992; Dorans & Lawrence, 1987; Dorans & Zeller, 2004a; Kulick & Hu, 1989; Schmitt & Bleistein, 1987). Most of

these studies have used either the standardization approach (Dorans & Kulick, 1986) and/or the Mantel–Haenszel procedures (Holland & Thayer, 1988) to study DIF.

The relationship between item difficulty and DIF was brought to national attention in 2003 when the article "Correcting the SAT's Ethnic and Social-Class Bias: A Method for Reestimating SAT Scores" by Roy Freedle was published in *The Harvard Educational Review* and then reached a broader audience through an article in the *Atlantic Monthly* that highlighted Freedle's findings (Mathews, 2003). Freedle's work described how many of the more difficult SAT items exhibited DIF benefiting African American students whereas easier SAT items showed DIF favoring White students. Freedle, a cognitive psychologist who worked at Educational Testing Service (ETS) for more than 30 years, hypothesized that the relationship he observed between item difficulty and DIF estimates was explained by cultural familiarity and semantic ambiguity. "Easy" Verbal items, he reasoned, tap into a more culturally specific content and therefore are hypothesized to be perceived differently, depending on one's particular cultural and socioeconomic background. "Hard" Verbal items, his hypothesis continued, often involve rarely used words that are less likely to have differences in interpretation across ethnic communities because these are only familiar to those with higher levels of education, which is more uniform than the home environment.[2]

Freedle (2003) reported a correlation of approximately 0.50 between the DIF statistics he used and the difficulty of the items. The empirical evidence on DIF he presented was obtained using the standardization approach. This statistical procedure was developed by Dorans and Kulick (1983). Freedle was criticized by Dorans and colleagues (Dorans, 2004; Dorans & Zeller, 2004a) and other researchers (Bridgeman & Burton, 2005; Wainer, 2009; Wainer & Skorupski, 2005), and his findings were attributed to the way Freedle applied the standardization approach, to the limitations of the standardization procedure, and to the role of guessing (among other issues).[3] When implementing the standardization approach appropriately, Dorans and Zeller (2004a) reported correlations somewhat smaller: a correlation of 0.27 between equated deltas and Mantel–Haenszel D-DIF and a correlation of 0.19 between equated deltas and the standardized formula score statistics for the African American/White comparison of Verbal items. They also found correlations of 0.31 and 0.22 for Analogy items.

The correlations described in Freedle's article were similar in magnitude to those described by earlier work on this area. Kulick and Hu (1989) reported a correlation between equated deltas and DIF estimates using the Mantel–Haenszel procedure for the White/African American comparison of 0.40 for the overall Verbal test and of 0.57 and 0.40 for the analogies and antonyms sections, respectively. Burton and Burton (1993) reported a correlation of 0.58 between Mantel–Haenszel D-DIF and item difficulty for a White/African American comparison conducted on 607 Analogy items from the 1987-1988 Verbal pretest item pool.

More recently, Scherbaum and Goldstein (2008) reported a correlation of –0.56 between item difficulty and DIF estimates obtained using the standardization approach in a large-scale civic knowledge test for ninth graders for the White/African American comparison (IEA Civic Education Study, 1999). The negative sign in this case indicates

that as proportion correct decreases (or item difficulty increases), DIF in favor of the focal group (African American students) increases. The correlation is 0.61 when a 2PL model is used to estimate item parameters and the Mantel–Haenszel is used to estimate DIF (Scherbaum & Goldstein, 2008). Santelices and Wilson (2010a) investigated the methodological concerns voiced by researchers about the way Freedle applied the standardization approach (Dorans, 2004, 2010) and found that, when the problems were indeed addressed, many of Freedle's claims remained steady. Although not as strong nor as widely spread as Freedle asserted, the authors found evidence supporting Freedle's original claim. They showed that African American students perform differentially more poorly in easier Verbal items when compared with White students, and differentially better than White students in harder Verbal items.

The current article deals with the hypotheses that attribute Freedle's findings to the limitations of the standardization approach. Dorans and Zeller (2004a) say,

> There is a strong possibility that this relationship [between item difficulty and DIF] is a statistical artifact because of the fact that large group differences in test-score distributions are not completely dealt with by the standardization, Mantel–Haenszel or, perhaps, any observed-score DIF procedure, and that impact remains in the data even after adjusting for total score differences. Since impact is positively related to difficulty, any DIF remaining in the impact may also be related to difficulty. There is a methodological issue here that needs to be addressed before we speculate about why the relationship exists. (p. 33)

Item Response Theory and DIF Analysis[4]

IRT provides a natural framework to study DIF since it directly models item responses. The essence of unidimensional item response models is that the probability of answering an item correctly or of attaining a particular response level is modeled as a function of an individual's ability and item difficulty parameters (Embretson & Reise, 2000). As the ability rises, the probability of answering an item correctly rises as well. The most salient representation of this relationship is the item characteristic curve (ICC), a mathematical function that provides the probability of a correct response to an item as a function of the person and item parameters. Under IRT, the examinee ability is not a direct linear transformation of number-correct test score; it is estimated taking into account parameters that describe the characteristics of the test items (Rogers, 1999).

The three most popular unidimensional IRT models for dichotomous item response data are the 1PL, 2PL, and 3PL models. In this framework, the analysis of DIF involves the matching of students on their ability level, which is a latent variable.

In an IRT perspective, DIF can be studied in a number of different ways provided that the IRT model fits the data and that there is sufficient sample size available. Three IRT methodologies to study DIF are the following: (a) parameter comparison (Brennan, 2006; Camilli & Shepard, 1994; Hambleton et al., 1991; Lord, 1980), (b) area

comparison (Brennan, 2006; Hambleton et al., 1991; Lord, 1980; Millsap & Everson, 1993), and (c) likelihood comparison (Holland & Wainer, 1993; Long, 1997; Thissen, Steinberg, & Wainer, 1988, 1993).

The Guessing Issue

Guessing has played an important role as a proposed explanation for the phenomenon Freedle described. Several investigations referred to guessing as a potential cause of the relationship between item difficulty and DIF estimates even before his 2003 article was published (Kulick & Hu, 1989; Schmitt & Bleistein, 1987). These authors explained the relationship between item difficulty and DIF estimates by pointing out the relative advantage on difficult items for the group that guesses differentially more: in their case, the White student group.

Bridgeman and Burton (2005) appealed to guessing again to explain Freedle's results. They argue that Freedle's findings can be explained as a statistical artifact based on the fact that in the standardization approach, students are matched on total scores rather than on true ability. Unlike true ability, they say, random responding can influence total scores.[5] According to them, guessing on questions that are far beyond students' skill level would be at the heart of the problem rather than a cultural/linguistic reason. Bridgeman and Burton (2005) illustrate their point through examples, data from questions without answer choices,[6] and responses from computer adaptive tests.

Guessing was also an important part of the argument made by the College Board when responding to Freedle's (2003) article. The response said,

> In brief, Freedle's suggestions boil down to capitalizing on chance performance. This kind of performance may represent either random guesses, or unconnected bits of knowledge that are not sufficiently organized to be of any use in college studies. (Camara & Sathy, 2004)

Wainer and colleagues (Wainer, 2009; Wainer & Skorupski, 2005) have also used guessing to explain the phenomenon described by Freedle. They 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 their argument is the assumption that, on average, White students have higher ability level than African American 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, they say, because of a "statistical artifact."

Guessing can occur on any type of test item, whether multiple choice or free response. However, it is generally considered to be a serious problem only on multiple-choice test results because of the greater likelihood of a correct response through guessing. The primary psychometric problem arising from guessing on test items is

that it increases the error variance of test scores, thereby reducing their reliability and validity (Rogers, 1999). If the IRT model does not explicitly take into account the noise that the guessing behavior adds to the regular problem-solving techniques, the estimation procedures may yield biased parameter estimates.

Traditionally, the research on guessing has focused on correction formulas applied to test scores after testing, and on the appropriateness of those corrections (Rogers, 1999). Although supporters of corrected scores stress the increased validity and reliability of corrected scores over noncorrected scores, detractors argue that examinees seldom guess at random and stress the confounding of personality traits with test scores (Rao & Sinharay, 2007; Yamamoto, 1987).

Within IRT some researchers have decided to model guessing as an item characteristic (Lord, 1980; Waller, 1989), whereas others have preferred to model it as a person parameter (Mislevy & Verhelst, 1990; Yamamoto, 1987). This latter modeling approach is based on the observation that the strategy used to answer multiple-choice questions is influenced by a personal decision being a student's choice to guess or not to guess (Xie, 2005).

Among the IRT models, we selected the 3PL model to address the issue of guessing because this correction-for-guessing-IRT model is best known in the educational measurement community. In addition, because of the complexities associated with estimating the *c* parameter in just one examinee population, DIF in the *c* parameter (which involves estimating the parameter in two populations) is not frequently studied. Our investigation will contribute to this less studied topic.

Research Questions

The research presented in this article explored the following questions:

- Is the relationship between item difficulty and DIF estimates observed for the White/African American comparison when methods that better control for different students' ability level (IRT methods) are used in the estimation of DIF?
- Is the relationship between item difficulty and DIF estimates still observed when the possibility of guessing is incorporated into the IRT model used to estimate DIF?

Data Sources and Sample

We studied the response patterns of all California students from public schools who took two specific SAT I forms in 1994 (Forms QI and DX) and two specific SAT I forms in 1999 (Forms IZ and VD). Note that this addresses one of the criticisms made by ETS researchers to Freedle's work: that the data he used were from old forms of the SAT and, therefore, preceded the procedures of DIF elimination implemented at ETS after 1980 (Dorans, 2004). The two forms from each year contain the same items but in different order. The four forms were chosen by the College Board to reflect the

usual variation between the versions of the SAT normally used in a given year and between years.

The test forms are more current than those analyzed by Freedle and were prescreened for DIF during item field tests. The files also contained the SAT Verbal and Math scores as well as students' responses to the student data questionnaire, which provides self-reported demographic and academic information such as parents' education, family income, and high school grade point average. Most of the information in the files comes from high school juniors. The sample used in this study included only high school juniors who reported speaking English as their best language. Focusing only on students whose best language is English allows isolating results from the effects of insufficient language proficiency in the population studied (Schmitt & Dorans, 1988).[7] The Hispanic group included students who reported being Latin American, Mexican, or Puerto Rican when asked about their ethnicity in the student data questionnaire.

Method

One-Parameter Logistic Model

Within the IRT framework, the parallel to the standardization approach is the 1PL model. The differences in the location of empirical ICCs studied by the standardization approach are conceptually parallel to the differences in item location, or "b parameter," of parametric ICCs modeled by IRT. Rather than matching examinees on observed score as it is done in the standardization method, the IRT methodology matches students on estimated ability.

To test Freedle's hypothesis of uniform DIF[8] using the IRT framework, we start by fitting the 1PL model to the data and then studying the relationship between DIF estimates and item difficulty when this model is used.[9]

Holland and Thayer (1986) showed that DIF analysis using the Rasch model to estimate differences in item difficulty parameters and the Mantel–Haenszel procedure produce identical results when the following conditions are met: (a) The matching variable is unbiased, (b) there is DIF on the studied item, (c) the focal and reference groups are random samples from their populations, and (d) the matching criteria include the studied item. Our study will allow us to first explore whether there are also similar results between the standardization approach and the Rasch model, to then move on to the 3PL model, a more complex IRT model.

The item characteristics curves for the 1PL model are given by the equation

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}}, \quad i = 1, 2, \ldots, I,$$

where $P_i(\theta)$ = the probability that a randomly chosen examinee with ability θ answers item i correctly, b_i = item i difficulty parameter, I = number of items in the test, and e = base of natural logarithm whose value is approximately 2.718.

$P_i(\theta)$ is an S-shaped curve with values between 0 and 1 over the ability scale. The b_i parameter for an item is the point on the ability scale where the probability of a correct response is .5. This parameter is a location parameter, indicating the position of the ICC in relation to the ability scale. The greater the value of the b_i parameter, the greater the ability that is required for an examinee to have a 50% chance of getting the item right; hence the harder the item (Hambleton et al., 1991).

One-Parameter Logistic Model and DIF as Implemented in ConQuest. The 1PL model was estimated using specialized IRT software called ConQuest (Wu, Adams, & Wilson, 1998). This software combines an item response model and a multivariate regression model (latent regression model); therefore, it is capable of estimating the item difficulty parameters controlling for the effect of ethnicity as well as ability level. This is required to adequately control for differences in mean group abilities, or impact.

The general form of the item response model fitted by ConQuest is the multidimensional random coefficient multinomial logit model (MRCML) described by Adams, Wilson, and Wang (1997). This model is flexible enough to allow the estimation of different Rasch-type IRT models,[10] including the random coefficient multinomial logit (RCML) within-logit-mean DIF model. The RCML within-logit-mean DIF model has the strength of allowing one to directly estimate DIF effect sizes and their standard errors as well as allowing the implementation of statistical tests of DIF indices (Moore, 1996; Paek, 2002). The MRCML and RCML within-logit-mean DIF model parameters are estimated by ConQuest using the marginal maximum likelihood method with the expectation maximization algorithm (Wu et al., 1998).

When modeling DIF using the RCML within-logit-mean DIF model, subgroups of test takers are matched on ability estimates provided by the 1PLM. In addition, a separate parameter is introduced to account for the variable of interest (ethnicity in this case). An item is deemed to exhibit DIF if the response probabilities for that item cannot be fully explained by the ability of the student and a fixed set of difficulty parameters for that item (Adams & Wilson, 1996; Wu et al., 1998). To model DIF, a parameter is introduced to account for the interaction of each item with ethnicity. Formula 1 depicts the RCML within-logit-mean DIF model

$$\text{logit}\{P_{ig}(\theta^*)\} = \theta^* - \delta_i + \Delta_g + \gamma_i G, \tag{1}$$

where $P_{ig}(\theta^*)$ refers to the probability of the response of a person in group g to item i; θ^* represents the examinee's ability distribution, which is independently and identically distributed. $N(\mu, \sigma)$; δ_i is the difficulty parameter for item i; γ_i refers to the DIF index parameter for item i; g indicates either the reference group or focal group that is compared with the reference group, $G = 1$ if $g = \text{R}$ (reference group), $G = 0$ if $g = \text{F}$ (focal group); and Δ_g is the mean ability difference between people in the reference and the focal group. For more details, see Paek (2002).

Three-Parameter Logistic Model

The ICCs for the 3PL model are described by the equation

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}, \quad i = 1, 2, \ldots, I,$$

where $P_i(\theta)$ and b_i are defined as in the Rasch model. The factor D is a scaling factor introduced to make the logistic function as close as possible to the normal ogive function. It typically assumes the value of 1.7. The parameter a_i, called the item discrimination parameter, is proportional to the slope of the ICC at the point b_i on the ability scale. At the point b_i, items with steeper slopes are more useful for separating examinees into different ability levels than are items with less steep slopes. These parameters are characteristic of the 2PL models.

The additional parameter in the 3PL model, c_i, is called the pseudoguessing parameter. This parameter provides a nonzero lower asymptote for the ICC and represents the probability of examinees with low ability answering the item correctly. It is important to note that by definition the c parameter is an item parameter and not a person parameter, therefore, it does not vary as a function of examinees' ability level. Also, the lowest and highest ability examinees have the same probability of getting the item correct by guessing.

Statistical Analyses

RCML within-logit-mean DIF model. The statistical significance of DIF was investigated using the likelihood-ratio test. The log likelihood-ratio test statistic is given by

$$-2\log\left(\frac{L_0}{L_1}\right) = -2[\log(L_0) - \log(L_1)] = -2(L_0 - L_1),$$

where L_0 refers to the log likelihood of the null model and L_1 refers to the log likelihood of the alternative model.

Under the null hypothesis of no DIF effects on item locations, the difference in these log likelihoods is distributed in large samples as χ^2 with $(n - 1)(m - 1)$ degrees of freedom, where n is the number of items and m is the number of groups. When this χ^2 is significant there is evidence that differential item effects are present (Long, 1997).

Models were also compared using the Bayesian information criterion (BIC, also known as Schwarz criterion) and the Akaike information criteria (AIC). The AIC is given by the following formula, where k is the number of parameters in the model:

$$\text{AIC} = -2\log(L) + 2(k + 1).$$

The BIC is given by

$$\text{BIC} = -2\log(L) + \ln(n)(k + 1),$$

where k is the number of parameters in the models and n is the number of observations in the sample. The AIC and BIC statistics provide two different ways of adjusting the

−2 log likelihood statistic for the number of terms in the model and the number of observations used. The three statistics described above are recommended for comparing nested models for the same data; the last two are also used to compare nonnested models; lower values of the statistic indicate a more desirable model (Long, 1997).

The effect size of any DIF effect found to be statistically significant was studied through the direct analysis of the interaction parameters of item difficulty and race/ethnicity estimated by ConQuest. This approach is similar to studying DIF through the difference in parameters of ICCs of different groups. The statistical significance of these DIF estimates was examined using the Wald test. The Wald test compares the maximum likelihood estimate of the parameter(s) of interest $\hat{\omega}$ to the proposed value ω with the assumption that the difference between the two will be approximately normal. The Wald statistic has a chi-square distribution with one degree of freedom (Agresti & Finlay, 1997). In the univariate case, the Wald statistics is

$$\text{Wald} = \frac{(\hat{\omega} - \omega)}{\text{Var}(\hat{\omega})}.$$

The effect size of the DIF estimates was analyzed following the guidelines provided by Paek (2002). Paek's guidelines are based on ETS effect size classification rules for the Mantel–Haenszel D-DIF statistic (Longford, Holland, & Thayer, 1993), which were reexpressed into logits using the statistical relationship between the Mantel–Haenszel D-DIF statistic and the 1PLM described by Holland and Thayer (1988). For the specific conditions under which these classification rules are equivalent, please refer to Paek (2002). The rules were derived to be compatible with the RCML DIF model Paek used; therefore, they had to be minimally adjusted[11] to conform with the parameterization of DIF in the RCML within-logit-mean model used in this section of the study. The effect size considerations used are the following:

$$|\text{DIF}| < 0.213: \text{Negligible}$$
$$0.213 \le |\text{DIF}| < 0.319: \text{Intermediate}$$
$$0.319 \le |\text{DIF}|: \text{Large}.$$

The RCML within-logit-mean model provides a set of item parameters for each group of examinees analyzed. In this case, there are two groups analyzed (with African American as the focal group and White students as the reference group); therefore, ConQuest was used to provide a set of item difficulty common to the reference group and the focal group, as well as a set of DIF estimates, one per item for each ethnicity. When examining the results it is necessary to keep in mind that the DIF estimates for the reference group are the negative of DIF estimates for the focal group. Note also that because of the way the RCML within-logit-mean model defines item difficulty parameters (a higher ability level is needed to reach the more difficult items) and DIF parameters (difference between the difficulty of the item in a given group and the average item difficulty), a negative DIF estimate indicates an item that is relatively

easier for the focal group under analysis while a positive DIF estimate indicates an item of differential higher difficulty for that same group.

The relationship between item difficulty and DIF was analyzed by directly inspecting the DIF estimates for items of different types and difficulty levels and by calculating correlations between item difficulty and DIF estimates. Positive correlations between item difficulty and DIF estimates among White students would confirm the relationship observed by Freedle since more difficult items would exhibit larger positive DIF estimates, indicating items that benefit the focal group, and easier items would exhibit small positive or negative DIF estimates, indicating items that benefit the reference group.

DIF Using the 3PL Model. The final deviance of the 1PL, 2PL, and 3PL models, as well as the final deviance of models that allow for uniform DIF in the 1PL model and 3PL model in all items simultaneously, were estimated using BILOG_MG (Scientific Software International, 2003).[12] The overall fit of these five models was studied using the likelihood-ratio test, the BIC, and the AIC.

Since software capabilities did not allow estimation of DIF for all three parameters of the 3PL model simultaneously, different item parameters for the focal and reference groups were obtained using a likelihood ratio test for DIF one item at the time (Thissen et al., 1988). To do this, the 3PL model was estimated in BILOG_MG using an all-but-the-studied-item anchor (Camilli & Shepard, 1994). Two models are estimated in this test, a constrained model in which no DIF is assumed and an unconstrained model in which DIF is assumed in the studied item. In the constrained model, a single set of a, b, and c parameters is estimated for the item, assuming that the ICC for both the reference and focal groups are the same. In the unconstrained model, all three parameters for the item under study are allowed to vary, yielding ICCs that potentially differ in difficulty, discrimination, and guessing. The remaining items are constrained to be equal in the two groups. Thus, two sets of items are obtained—a_f, b_f, c_f and a_r, b_r, c_r. The unconstrained model estimates six parameters (as opposed to three for the constrained model) and the difference is tested with the log likelihood statistic. This statistic is distributed with three degrees of freedom, corresponding to the additional three parameters. Although the same constrained model is used for all items, a different unconstrained model is estimated for each studied item (A. Cohen & Bolt, 2005).

In a sense, the DIF estimate obtained by Freedle for each item using the standardization procedure here is parallel to the differences between all three of the parameters estimated in the focal and the reference groups ($\Delta a = a_r - a_f$, $\Delta b = b_r - b_f$, $\Delta c = c_r - c_f$). In our case, we were most interested in the difference between parameters b and c and their effect size. The study of differences in parameter b would allow us to explore Freedle's phenomenon in the 3PL model and compare the results with those obtained from the Rasch model, which only allows for differences in parameter b. In addition, differences in parameter c aim to examine differences between the focal and the reference groups in the probability of low-ability students answering items correctly by chance, which is one of the hypotheses offered by researchers for the

Freedle phenomenon (Wainer, 2009; Wainer & Skorupski, 2005). We did not investigate deeply differences in the parameter *a* as they were not readily interpretable and no accepted criteria to classify the size of the differences were found in the literature.

To compare results with those obtained from the Rasch model, the difference between parameters *b* (Δb) is classified using the cutoff scores presented in the previous section which are themselves based on ETS effect size classification rules for the Mantel–Haenszel D-DIF statistic. However, since DIF in the 3PL model context is given by the difference between the parameters estimated for each group, and not by the difference between the group parameter and the parameter average as in RCML within-logit-mean model, the cutoff scores needed to be adjusted (effectively, they are doubled). Thus, the guidelines used in this section of the study are the following:

$$|\text{DIF}| < 0.426: \text{Negligible}$$
$$0.426 \leq |\text{DIF}| < 0.638: \text{Intermediate}$$
$$0.638 \leq |\text{DIF}|: \text{Large}.$$

The difference between parameters *c* (Δc) was classified using the cutoff scores used in a simulation study by Thissen et al. (1988).[13] These cutoff scores were expressed in numeric terms and no flag for test development was attached to them. The cutoff scores are the following:

$$|\text{DIF}| < 0.05$$
$$0.05 \leq |\text{DIF}| < 0.10$$
$$0.10 \leq |\text{DIF}| < 0.15$$
$$|\text{DIF}| \geq 0.15.$$

The relationship between the item difficulty parameter and the difference between item parameters was explored using correlation analysis. A positive correlation between item difficulty and Δb in the 3PL model, which explicitly considers the probability of low-ability students responding correctly because of guessing, would provide evidence in support of Freedle's claim.

Since only a weak correlation was found between the item difficulty parameters and the DIF estimates in the Math test when using both the standardization approach and the Rasch within-mean-logits model, the Freedle phenomenon is only explored in the Verbal test.

Results

We present first the results from the analyses using the 1PL model first and then the results from the analyses using the 3PL model.

Table 1. Model Comparison Statistics (Verbal Test Form IZ 1999)

Model	−2 Log likelihood	BIC	AIC
1PL model	559,818.423	560,531.22	559,978.42
DIF model	559,040.837	560,439.69	559,354.84
Difference	777.586	91.53	623.58

Note. 1PL model = one-parameter logistic model; BIC = Bayesian information criterion; AIC = Akaike information criterion.

RCML Within-Logit-Mean DIF Model

This section presents the results from two sets of analyses. Initially, the results from the Verbal test of the SAT Form IZ are presented in detail. The reference group comprised White students and the focal group comprised African American students. Each of the SAT tests was treated as unidimensional. Form IZ was chosen for initial exploratory analyses because it was one of the most recent forms available and had a large focal group.

The second set of analyses explores the generalizability of the first set of results across ethnic groups and test forms. Freedle claimed that the linguistic ambiguity explanations would hold across languages and ethnic groups. To shed light into this issue, DIF was analyzed using the 1PL model among White and African American students first, and then among White and Hispanic students, in four SAT forms (DX, QI, IZ, and VD).

DIF in the SAT Form IZ. The results from the Verbal test of Form IZ, which was administered to 7,405 students from California public high schools in 1999 (6,548 White students and 857 African American students), replicate the relationship described by Freedle (2003) and Santelices and Wilson (2010b).[14]

Initially, the fit of two models was compared for the Verbal test through the likelihood ratio test comparisons: (a) the 1PL model and (b) the Rasch within-logit-mean DIF model (DIF model). This likelihood ratio statistic is distributed as a chi-square distribution with 77 degrees of freedom (78 items and 2 groups) and the value shown in Table 1 is statistically significant at a .005 confidence level suggesting that the DIF model fits the data from the Verbal test better than the 1PLmodel. The Schwarz criterion and AIC also support the better fit of the DIF model.

The Wald test identifies 61 items with statistically significant DIF. In total, 50 of the 61 items exhibit negligible DIF, 8 show intermediate DIF, and 3 have large DIF estimates.

Figure 1 shows the relationship between item difficulty and DIF estimates[15] obtained when fitting the DIF model in ConQuest. The correlation between the item difficulty and the DIF estimates is $r = .62$ (statistically significant at the $\alpha = .001$ level). In Figure 1, the items have been ordered from easiest (left) to hardest (right) using item difficulty estimates from the reference group; the diagonal solid line shows the line of best fit observed in each graph's data (using a linear regression estimate). The results replicate the phenomenon described by Freedle (2003) and Santelices and

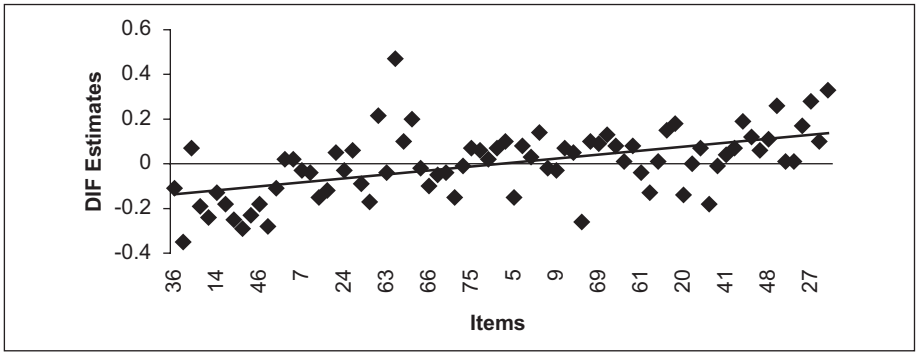

Figure 1. Relationship between item difficulty and differential item functioning (DIF) estimates

Note. Item difficulty and DIF estimates were obtained from the Rasch within-logit-mean differential item functioning model in the Verbal test of Form IZ 1999. Items are ordered by the difficulty level estimated for the reference group.

Wilson (2010b): Easier items exhibit DIF in favor of the White group (negative DIF estimates) whereas harder items show DIF favoring the African American group (positive DIF estimates). This correlation between DIF estimates and item difficulty estimates is consistent with Freedle's results: He obtained correlations of 0.52, 0.41, and 0.48 for the Analogies, Antonyms, and Sentence Completion items, respectively. Only the Reading Comprehension items showed a lower and nonsignificant correlation of 0.08. The correlation observed here is also somewhat larger than the correlations obtained when replicating Freedle's methodology (the standardization approach) using more recent data (in Santelices & Wilson, 2010b, the correlation was between 0.36 and 0.42).

Not only do we observe the general phenomenon Freedle described in the overall correlation but also when analyzing DIF estimates by different item types. Furthermore, and despite the fact that Freedle did not find the pattern in reading comprehension items, we observe results in the general direction he described in all three types of Verbal items: Analogies (r = .59), Reading Comprehension items (r = .64), and Sentence Completion items (r = .61; see Table A1 in the appendix for more details).[16]

Relationship between DIF and item difficulty across test forms and ethnic groups. This section reports the results obtained when exploring Freedle's phenomenon in four test forms (DX, QI, IZ, and VD) and across two ethnic groups. The test forms DX and QI were administered in 1994 and forms IZ and VD were administered in 1999. DIF was first analyzed among White and African American students and then among White and Hispanic students. All sample sizes were adequate to conduct DIF analyses (Clauser & Mazor, 1998). See Table 2 for details on the size of each group.

Because of the low correlation between item difficulty and DIF estimates found in the Math test when using the 1PL model and the standardization approach to DIF and since most research shows only evidence of small and less systematic DIF in the Math test than in the Verbal test (Kulick & Hu, 1989; O'Neill & McPeek, 1993; Schmitt &

Table 2. Sample Sizes

Group	1999 IZ	1999 VD	1994 QI	1994 DX
White students	6,548	6,682	3,360	3,188
Hispanic students	1,904	2,018	982	1,003
African American students	857	929	671	709

Table 3. Correlation Between Differential Item Functioning (DIF) Estimates and Item Difficulty Across Ethnic Groups and Test Forms Using the Rasch Within-Logit-Mean DIF Model in the Verbal Tests

Ethnic group	Correlation	1999 IZ	1999 VD	1994 QI	1994 DX
White, African American	Parameter *b*, DIF estimates	.621	–.152	.426	.463
	Prob > |*r*| under H_0: $Rh_o = 0$	<.0001	.1854	.0001	<.0001
White, Hispanics	Parameter *b*, DIF estimates	.439	.373	.156	.219
	Prob > |*r*| under H_0: $Rh_o = 0$	<.0001	.0008	.1724	.0546

Dorans, 1988; Zwick, 2002), this section focuses only on the Verbal test of forms IZ, VD, QI, and DX.[17]

The results presented in Table 3 replicate the relationship between item difficulty and DIF estimates in test forms more current that those analyzed by Freedle and when using methods that better control for the difference in groups' mean ability. This relationship is somewhat stronger for the White/African American comparison but it is also observed when comparing item functioning between White and Hispanic students. Three of the four forms analyzed present correlations in the range observed by Freedle for the White/African American comparison and all three are statistically significant. Although correlations are lower for the White/Hispanic comparison, two of them are greater than 0.35 and statistically significant.

The results for the White/African American comparison are similar to those obtained when applying the standardization method to the same four forms (Santelices & Wilson, 2010b). The relationship described by Freedle is observed in the same three forms under both methodologies (Forms IZ, QI, and DX); Form VD does not exhibit this relationship under any of the methodologies. The correlations are larger, however, when modeling DIF using the Rasch within-mean-logits model than when using the standardization approach.

On the other hand, the results for the White/Hispanic comparison are different under the two methodologies (see Table 4). Although none of the forms presented strong and statistically significant correlations when using the standardization approach, the most recent two forms (IZ and VD) present evidence supporting the relationship between item difficulty and DIF estimates when using the Rasch within-mean-logits DIF model. The correlation between item difficulty and DIF estimates observed in these two forms is in the range reported by Freedle.

The results presented in this section suggest that DIF methods that better control for the difference in groups' mean ability level find even stronger evidence in support of Freedle results.

Table 4. Correlation Between *p* Value and DIF Estimates Across Ethnic Groups and Test Forms Using the Standardization Approach in Verbal Tests

		Form			
Group	DIF method	1999 IZ	1999 VD	1994 QI	1994 DX
White, African American	Prop Correct, STD P DIF	–.414***	–.141	–.317**	–.257*
	Prop Correct, STD FS DIF	–.420***	–.166	–.293**	–.240*
White, Hispanics	Prop Correct, STD P DIF	–.179	–.101	–.009	.038
	Prop Correct, STD FS DIF	–.182	–.084	.020	.038

Note. From Santelices and Wilson (2010b). Reprinted with permission. Copyright 2010 by the President and Fellows of Harvard College. For more information, please visit http://www.harvardeducationalreview.org

*$p < .05$. **$p < .01$. ***$p < .001$.

DIF Using the 3PL Model

This section presents the results from two sets of analyses. Again, the results from the Verbal test of the SAT Form IZ are initially presented in detail. For this analysis the reference group consisted of White students and the focal group consisted of African American students. The second set of analyses explores the generalizability of the first set of results across ethnic groups and forms. Freedle claimed that the linguistic ambiguity explanations hold across languages and ethnic groups. To shed light on this issue, DIF was analyzed using the 3PL model among White and African American students first, and then among White and Hispanic students, in four SAT forms (DX, QI, IZ, and VD).

DIF in the SAT Form IZ. The model fit of several models was compared through the likelihood ratio test, the BIC, and the AIC. Because of limitations in the estimation capacity of the software used, these statistics were only available for models that allowed for uniform DIF. The results displayed in Table 5 show that when modeling the data with the 3PLM, which includes the pseudo-guessing parameter, there is still evidence of uniform DIF.

In addition, the differences between parameters b (Δb) and between parameters c (Δc) are displayed in Tables 6 and 7 classified by size. Although only 2 of the 78 items that classify for the largest categories (3%) show differences in the c parameter, 9 of the 78 items (12%) exhibit large DIF in the b-parameter.[18] These tables suggest that most items exhibit small DIF in the b and c parameters.

In general, the results from the correlation analysis tend to support the phenomenon described by Freedle. Table 8 shows the correlations between item difficulty and the differences in the item parameters. The largest correlation is observed between item difficulty and Δb, just as in Freedle's analysis: Easier items tend to benefit White students whereas more difficult items benefit African American students. Furthermore, the strength of the correlation is similar in magnitude to the correlation observed by Freedle and other researchers. The correlation to the difference in the pseudo-guessing

Table 5. Statistics of Model Fit (Verbal Test Form IZ 1999)

Model	–2 Log likelihood	BIC	AIC
3PLM	608,389.4	610,474.32	608,857.40
3PLM_DIF_b[a]	607,075.4	609,855.29	607,699.40
Difference	1,314.00	619.03	1,158.00

Note. BIC = Bayesian information criterion; AIC = Akaike information criteria; 3PLM = three-parameter logistic model.
a. Item discrimination and item guessing parameters are the same for the focal and reference groups when uniform DIF is analyzed using the 3PLM. These parameters are, however, allowed to vary from item to item.

Table 6. Distribution of Δb From the 3PL Model Log Likelihood Ratio Test, All-but-the-Studied-Item Anchor (Verbal Test Form IZ 1999)

Difference in parameter *b*	Number of Items
Negligible DIF or |DIF| < 0.426	58
Intermediate DIF or 0.426 < |DIF| < 0.638	11
Large DIF or |DIF| ≥ 0.638	9
Total	78

Note. 3PL model = three-parameter logistic model; DIF = differential item functioning.

Table 7. Distribution of Δc From the 3PL Model Log Likelihood Ratio Test, All-but-the-Studied-Item Anchor (Verbal Test Form IZ 1999)

Difference in parameter *c*	Number of Items
[Diff. > –0.05, Diff. < 0.05]	57
[–0.10 < Diff. < –0.05, 0.05 < Diff. < 0.10]	15
[–0.15 < Diff. < –0.10, 0.10 < Diff. < 0.15]	4
[Diff. < –0.15, Diff. > 0.15]	2
Total	78

Note. 3PL model = three-parameter logistic model; Diff. = difference.

parameter (Δc) is small (J. Cohen, 1969/1988) and statistically nonsignificant; therefore, item difficulty does not tend to be related to differences in item guessing between the groups. The correlation to the difference in the *a* parameter (discrimination parameter) is moderate and statistically significant.

The graphical relationship between item difficulty and the difference of item difficulty is shown in Figure 2. The solid line represents the line of best fit observed. Although more difficult items tend to benefit African American students, easier items tend to benefit White students.

DIF across forms and ethnic groups. This section reports the results obtained when exploring Freedle's phenomenon in the Verbal test of four SAT forms (DX, QI, IZ, and VD) and in two ethnic groups using an IRT model that takes into account the possibility

Table 8. Correlation Between Item Difficulty and DIF Estimates When Using the 3PL Model (Verbal Test of Form IZ 1999)

Item difficulty (b) compared with	Correlation	*p* Value
Δa	0.297	.0083
Δb	0.518	<.0001
Δc	0.069	.550

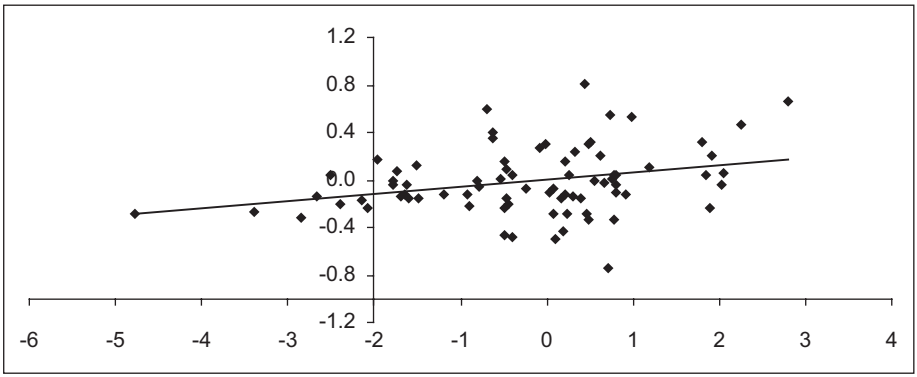

Figure 2. Relationship between item difficulty and differential item functioning (DIF) estimates in the item difficulty parameter
Note. Scatter plot of item difficulty (*X*) and difference in item difficulty (*Y*) when using the three-parameter logistic model in the Verbal test of Form IZ 1999.

of students answering items correctly by guessing. The test forms DX and QI were administered in 1994 and forms IZ and VD were administered in 1999. DIF was first analyzed among White and African American students and then among White and Hispanic students. All sample sizes were adequate to conduct DIF analyses (Clauser & Mazor, 1998).

Software limitations did not allow us to model fit allowing for DIF in all parameters simultaneously. Alternatively model fit was explored using models that allowed for uniform DIF (DIF only in parameter *b*). The results displayed in Table 9 show that when modeling the data with the 3PL model, which considers the possibility of guessing, there is still evidence of uniform DIF across all forms and ethnic groups. The 3PLM that allows for different parameters *b* in the focal and reference groups (3PLM_DIF) shows smaller deviance than any of the other three models under study (1PLM, DIF in the 1PLM, and 3PLM) in all four test forms and both when comparing White examinees with African American and to Hispanic examinees. The 3PL model that allows for different parameters *b* in the focal and reference groups (3PLM_DIF) shows smaller log likelihoods than the 3PL model in all four test forms and both when comparing White examinees with African American and with Hispanic examinees. The differences in relative log likelihoods observed between the SAT forms administered in 1999 (Forms IZ and VD) and in 1994 (QI and DX) are explained, at least in part, by the difference in the number of examinees from each year.

Table 9. Model Fit (–2 Log Likelihood) Across Test Forms and Ethnic Groups (Verbal Tests)

Model	IZ 99	VD 99	QI 94	DX 94
White/African American comparison				
1PLM	614,594.4	632,141.67	355,716.02	344,220.31
1PLM_DIF	612,929.8	630,469.48	354,386.95	343,067.15
3PLM	608,389.4	625,874.08	351,510.13	340,207.89
3PLM_DIF_b[a]	607,075.4	624,594.50	350,640.96	339,444.27
Difference [3PLM-3PLM]	1,314.00	1,279.58	869.17	763.62
Sample size				
White students	6,548	6,682	3,360	3,188
African American students	857	929	671	709
White/Hispanic comparison				
1PLM	704,859.85	727,053.56	382,427.56	369,359.18
1PLM_DIF	703,216.99	725,255.54	381,767.16	368,648.48
3PLM	697,683.03	719,773.10	378,352.95	364,975.56
3PLM_DIF_b[a]	696,488.30	718,564.94	377,893.83	364,495.02
Difference [3PLM_DIF-3PLM]	1,194.73	1,208.16	459.12	480.54
Sample size				
White students	6,548	6,682	3,360	3,188
Hispanic students	1,904	2,018	982	1,003

Note. 1PLM = one-parameter logistic model; 3PLM = three-parameter logistic model; DIF = differential item functioning.

a. Item discrimination and item guessing parameters are the same for the focal and reference groups when uniform DIF is analyzed using the 3PLM. These parameters are allowed to vary, however, from item to item.

The overall DIF effect, which in the case of the standardization approach and the Rasch model was measured by just one parameter, in the 3PL model is given by the difference between the three-parameter estimates from the focal and the reference groups (Δa, Δb, Δc). To get DIF estimates for parameters a, b, and c simultaneously the log likelihood-ratio test for DIF was implemented using the all-but-the-studied item anchor. The analysis was conducted for one item at the time and, hence, it was not possible to obtain an overall model fit.

Tables 10 and 11 show the distribution of the difference in parameters b and c across forms and ethnic groups using the cutoff scores described in the "Statistical Analyses" subsection. The evidence suggests that the DIF phenomenon observed when modeling the data with the 3PL model, which takes guessing into account, is slightly more related to differences in parameter b than to differences in parameter c when comparing White and African American students. Looking across forms, for the White/African American comparison approximately 10% of the items exhibit large DIF, whereas about 4% of the items exhibit differences in parameter c that classify for the largest classification bin. There is not a clear pattern for the White/Hispanic comparison across forms: The percentage of items exhibiting large DIF in the item difficulty parameter ranges between 3% (Form QI 94) and 10% (Form IZ 99) and the

Table 10. Difference in Parameters *b* Across Test Forms and Ethnic Groups (3PL Model Log Likelihood Ratio Using the All-but-the-Studied-Item Anchor in Verbal Tests)

Form	Negligible DIF |DIF| < 0.426	Intermediate DIF 0.426 ≤ |DIF| < 0.638	Large DIF |DIF| ≥ 0.638	Percentage of items with large DIF	Total
White/African American comparison					
IZ 99	58	11	9	12	78
VD 99	64	6	8	10	78
QI 94	66	5	7	9	78
DX 94	64	8	6	8	78
White/Hispanic comparison					
IZ 99	63	7	8	10	78
VD 99	68	5	5	6	78
QI 94	71	5	2	3	78
DX 94	71	3	4	5	78

Note. 3PL model = three-parameter logistic model; DIF = differential item functioning.

percentage of items exhibiting large differences in the pseudoguessing parameter ranges between 4% (Forms VD 99 and DX 94) and 19% (Form QI 94).

Analyzing the relationship between the item difficulty parameters and each of the parameter differences allows us to better understand whether the phenomenon Freedle observed was because of DIF in the item difficulty parameter or of DIF in the pseudo-guessing parameter. Positive and statistically significant correlations between the item difficulty parameter and the difference in parameter *b* would replicate Freedle's results: Easier items tend to benefit White students whereas more difficult items tend to benefit African American students.

The results presented in Table 12 suggest that the Freedle phenomenon is indeed still observed when the IRT model includes the pseudo-guessing parameter. This is so even when we go beyond the usual 3PL model and allow the guessing rates to vary between ethnic groups. For the White/African American comparison we observe the phenomenon in three forms out of four (IZ, VD, and DX) and for the White/Hispanic comparison we observe it in two forms out of four (IZ and VD). The moderate (J. Cohen, 1969/1988) and statistically significant correlations observed between Δb and the item difficulty parameter (in three of the forms for the White/African American comparison and in two forms for the White/Hispanic comparison) and the low and nonsignificant correlations found between Δc and the item difficulty parameter support the idea that the phenomenon observed by Freedle is associated to DIF in the item difficulty parameter and not so much by DIF associated to the pseudoguessing parameter.

The results presented in this section suggest that IRT models that take guessing into consideration do not negate the empirical relationship observed by Freedle for the Verbal test: More difficult items continue to exhibit DIF in favor of African American students. Although there is variation in the strength of the relationship across forms

Table 11. Difference in Parameters *c* Across Test Forms and Ethnic Groups (3PL Model Log Likelihood Ratio Using the All-but-the-Studied-Item Anchor in Verbal Tests)

Form	[Diff. > –0.05, Diff. < 0.05]	[–0.10 < Diff. < –0.05, 0.05 < Diff. < 0.10]	[–0.15 < Diff. < –0.10, 0.10 < Diff. < 0.15]	[Diff. < –0.15, Diff. > 0.15]	Percentage of items with |DIF| > 0.15	Total
White/African American comparison						
IZ 99	57	15	4	2	3	78
VD 99	52	17	6	3	4	78
QI 94	59	13	3	3	4	78
DX 94	47	24	3	4	5	78
White/Hispanic comparison						
IZ 99	57	11	4	6	8	78
VD 99	56	14	5	3	4	78
QI 94	18	29	16	15	19	78
DX 94	59	12	4	3	4	78

Note. 3PL model = three-parameter logistic model; DIF = differential item functioning.

and ethnic groups, there is enough evidence to consider this phenomenon real and independent of guessing and methodological concerns.

Discussion of Results

This study investigated two hypotheses using an IRT approach:

1. The hypothesis that the relationship between item difficulty and DIF would be ameliorated when using methods that better control for differences between impact and DIF (in comparison with the standardization approach).
2. The hypothesis that the relationship between item difficulty and DIF would diminish when considering the role of guessing.

The IRT methodology allowed us to better control for differences in the mean ability level of the groups under study since parameter estimation is less sample-dependent than in observed-score DIF methodology. In addition, the explicitness and flexibility of the IRT models are convenient for exploring different hypotheses regarding the relationship between item difficulty and DIF estimates.

The results presented in this article suggest that DIF methods that aim to better control for the difference in groups' mean ability level, and in data sets more current than those analyzed by Freedle, generally find evidence in support of the relationship between item difficulty and DIF. This relationship is observed more frequently than when using the standardization approach.

The relationship between item difficulty and DIF estimates described by Freedle is observed when using the RCML within-logit-mean DIF model in the Verbal test across forms and ethnic groups: The "Freedle phenomenon" is observed in three forms out of

Table 12. Correlation Between DIF Estimates and Item Difficulty Across Ethnic Groups and Test Forms Using the Three-Parameter Logistic Model (Verbal Tests)

Group	Item Difficulty Correlated With	1999 IZ	1999 VD	1994 QI	1994 DX
White,	Δa	.297	.097	.174	.026
African	(Prob > |r| under H_0: $Rh_o = 0$)	(.0083)	(0.3946)	(.1272)	(.8239)
American	Δb	.518	.512	–.186	0.435
	(Prob > |r| under H_0: $Rh_o = 0$)	(<.0001)	(<.0001)	(.1039)	(<.0001)
	Δc	.069	.030	–.230	–.027
	(Prob > |r| under H_0: $Rh_o = 0$)	(.550)	(.7949)	(.0428)	(.8118)
White,	Δa	.335	.171	.0876	.120
Hispanics	(Prob > |r| under H_0: $Rh_o = 0$)	(.0027)	(.1309)	(.4459)	(.294)
	Δb	.452	.317	–.061	.179
	(Prob > |r| under H_0: $Rh_o = 0$)	(<.0001)	(.0044)	(.5986)	(.116)
	Δc	.070	.019	–.096	.012
	(Prob > |r| under H_0: $Rh_o = 0$)	(.5421)	(.8703)	(.4037)	(.9142)

four for the White/African American comparison (IZ, QI, and DX) and in two forms out of four for the White/Hispanic comparison (IZ and VD).

The relationship between item difficulty and DIF estimates is also observed when using the 3PL model, which includes the possibility that students respond correctly to an item because of random guessing. For the White/African American comparison we observe the phenomenon in three forms out of four (IZ, VD, and DX) and for the White/Hispanic comparison we observe it in two forms out of four (IZ and VD).

For the White/Hispanic comparison (see Table 13) the results from the analyses using the Rasch model and the 3PL model are quite consistent. Moderate correlations between item difficulty and DIF estimates (Δb in the 3PLM) are found in the same two forms under both methodologies (IZ and VD), although not for Forms QI and DX. Santelices and Wilson (2010b) found no evidence in support of the relationship when using the standardization approach; therefore, other IRT methods find evidence of the Freedle phenomenon for the White/Hispanic comparison more frequently than the standardization approach.

For the White/African American comparison, as in the case of the White/Hispanic comparison, the IRT methodology finds evidence of the "Freedle phenomenon" more frequently than the standardization approach. Although the standardization approach supports the relationship between item difficulty and DIF estimates in two of the four forms (Santelices & Wilson, 2010b), when using each IRT model the phenomenon is observed in three of the four forms. The IRT results for the White/African American comparison, however, are somewhat less consistent than those obtained for the White/Hispanic comparison. For the White/African American comparison we obtain similar results from the Rasch and the 3PL model in two of the four forms analyzed (1999 IZ

Table 13. Presence of the Freedle Phenomenon Across Methods, Forms, and Ethnic Groups (Verbal Tests)

Group	Method	1999 IZ	1999 VD	1994 QI	1994 DX
White, African American	Standardization approach	Yes	No	Yes	No
	Rasch model	Yes	No	Yes	Yes
	3PLM	Yes	Yes	No	Yes
White, Hispanics	Standardization approach	No	No	No	No
	Rasch model	Yes	Yes	No	No
	3PLM	Yes	Yes	No	No

Note. 3PLM = three-parameter logistic model. "Presence" of the Freedle phenomenon is defined as a statistically significant and moderate correlation (>0.3).

and 1994 DX). In the other two forms, however, the Freedle phenomenon is only observed under one of these two IRT models. The Freedle phenomenon is found in Form 1999 VD only when the 3PLM is used and in Form 1994 QI only when the Rasch model is used.

The correlations described in this article are similar in magnitude to those described by Freedle (2003) and other researchers. See, for example, Kulick and Hu (1989), Burton and Burton (1993), Scherbaum and Goldstein (2008), and Santelices and Wilson (2010b). At the same time, it is important to note that most items exhibited small or medium DIF estimates.

This study contributes to the literature by expanding the methods used to analyze DIF and its relationship with item difficulty. This relationship between DIF and item difficulty has been observed when DIF is studied using the Mantel–Haenszel procedure (Dorans & Zeller, 2004a; Kulick & Hu, 1989), the standardization approach (Freedle, 2003; Kulick & Hu, 1989; Santelices & Wilson, 2010b; Scherbaum & Goldstein, 2008), and now the two new IRT approaches presented in this article (i.e., this article has presented analyses using the Rasch model and the 3PL model whereas Scherbaum & Goldstein, 2008, used the 2PL model to study DIF).

Results from the article support the conclusion from Kulick and Hu (1989) when studying alternative explanations to the relationship between DIF and item difficulty: "In general, item difficulty is related to DIF. The nature of that relationship appears to be independent of the choice of DIF index (either the Mantel–Haenszel or the standardization approach) as well as of test form" (p. 1). Our choice of DIF indices was different from Kulick and Hu's (we used the Rasch model and the 3PL model) but our conclusion is the same: A relationship between item difficulty and DIF is observed in the Verbal test of the SAT, such that more difficult items tend to exhibit DIF in favor of the focal group (African American examinees) and easier items exhibit DIF in favor of the reference group (White examinees). Although the different methodologies analyzed do not provide completely consistent results, the methodological considerations at the

heart of this article are not enough to make the relationship between DIF and item difficulty disappear.

Conclusion

This investigation set out to explore two hypotheses offered by researchers as explanation of the empirical phenomenon Freedle described, namely (a) the limited capacity of observed-score DIF methodology to control for differences in the mean ability level of the groups compared (Dorans & Zeller, 2004a) and (b) the role of guessing in multiple-choice items (Camara & Sathy, 2004; Wainer, 2009; Wainer & Skorupski, 2005).

This study investigated both these hypotheses using IRT methods, which allow one to better control for differences in the mean ability level of the groups under study since parameter estimation is less sample dependent than in classical test theory methodology.

We used two different IRT models: the Rasch and the 3PL model. The Rasch model is the IRT equivalent of the standardization approach since items are characterized exclusively by their relative difficulty. The 3PL model, a more complex IRT model, explicitly incorporates the possibility that low-ability students respond to an item correctly by chance and also allows items to vary in the discrimination parameter.

We find evidence of a moderate relationship between item difficulty and DIF estimates in the Verbal test across ethnic groups and forms when using both these IRT models. When using the RCML within-logit-mean DIF model across Verbal forms the relationship between item difficulty and DIF estimates described by Freedle is supported in three forms out of four for the White/African American comparison and in two forms out of four for the White/Hispanic comparison. When using the 3PL model, the relationship between item difficulty and DIF estimates was observed in three forms out of four for the White/African American comparison and in two forms out of four for the White/Hispanic comparison.

In general, our results when using IRT methodology to estimate DIF and item parameters in data sets more current than those analyzed by Freedle showed evidence in support of the relationship between DIF and item difficulty and do not support the alternative hypothesis offered by other researchers regarding the role of guessing and the limitations of the standardization approach. None of these considerations were able to eliminate the relationship between DIF and item difficulty.

In our judgment, the replication of Freedle's findings when using both observed-score DIF methodology (Santelices & Wilson, 2010b, Scherbaum & Goldstein, 2008) and IRT methodology (this article and Scherbaum & Goldstein, 2008) provide evidence for the "Freedle phenomenon." The investigation of potential causes should include studies that investigate Freedle's proposed explanation, the influence of academic versus home language (Freedle, 2010) including investigation of the

Appendix

Table A1. Mean Differential Item Functioning (DIF) Estimates by Item Type Using the Rasch Within-Logit-Mean DIF Model in the Verbal Test of Form IZ 1999

Item type	Difficulty level	Mean difficulty	Mean DIF	Group who benefits	Statistically significant DIF (Wald test)[a]	Percentage by item group significant (Wald test)	Statistically significant DIF and |DIF| > 0.213	Percentage by item type significant DIF and |DIF| > 0.213	Total number of items
Analogies (13)	Easier 5	−1.33	−0.140	Reference	5	100	2	40	5
	Harder 5	0.99	0.028	Focal	5	100	0	0	5
Reading Comprehension (40)	Easier 5	−0.99	−0.118	Reference	3	60	2	40	5
	Middle 5	0.50	0.052	Focal	3	60	0	0	5
	Harder 5	1.31	0.108	Focal	3	60	1	20	5
Sentence Completion (25)	Easier 5	−1.70	−0.180	Reference	5	100	1	20	5
	Middle 5	0.56	−0.064	Reference	3	60	1	20	5
	Harder 5	1.76	0.162	Focal	4	80	2	40	5
Total number of items					31	78	9	23	40

a. Individual items are statistically significant at alpha level of .05.

Table A2. Mean Differential Item Functioning (DIF) Estimates by Item Type When Using the Rasch Within-Logit-Mean DIF Model in the Math Test of Form IZ 1999

Item type	Relative item difficulty	Mean difficulty	Mean DIF	Group benefits	Statistically significant DIF (Wald test)[a]	Percentage by item group significance (Wald test)	Statistical Significance and |DIF| > 0.213	Percentage by item type statistical significance and |DIF| > 0.213	Total number of items
Multiple Choice (35)	Easier 5	−1.84	−0.09	Reference	3	60	0	0	5
	Middle 5	−0.18	−0.12	Reference	4	80	2	40	5
	Harder 5	2.37	0.06	Focal	3	60	2	40	5
Quantitative Comparison (14)	Easier 5	−1.29	0.02	Focal	3	60	1	20	5
	Harder 5	0.86	0.03	Focal	3	60	0	0	5
Student Produced Response (10)	Easier 5	−0.95	−0.08	Reference	5	100	0	0	5
	Harder 5	0.91	0.01	Focal	4	80	2	40	5
Total number of items					25	71	7	20	35

a. Individual items are statistically significant at alpha level of .05.

Table A3. Correlation Between DIF Estimates and P-Values Using using the Rasch within-logits-mean DIF model in the Math Test Across Ethnic Groups and Test Forms

Group	Correlation Parameter b, DIF estimates	Form			
		1999 IZ	1999 VD	1994 QI	1994 DX
White, African Amer.	Parameter b, DIF estimates	0.204	0.182	0.379	0.363
	Prob > |r| under H_0: Rh_o=0	0.120	0.168	0.003	0.004
White, Hispanics	Parameter b, DIF estimates	0.204	0.125	0.329	0.408
	Prob > |r| under H_0: Rh_o=0	0.119	0.345	0.010	0.001

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 sensibility of Freedle's phenomenon to alternative forms of guessing such as differential guessing strategies between White and students from other ethnic groups.

Acknowledgments

We wish to thank Saul Geiser for helpful conversations and the College Board for providing the data.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was funded by UC ACCORD through a dissertation fellowship.

Notes

1. Kulick and Hu (1989) also implemented the Mantel–Haenszel procedure.
2. For more information regarding DIF in the SAT, see Carlton and Harris (1992), O'Neill and McPeek (1993), Schmitt and Bleistein (1987).
3. Researchers at ETS (Bridgeman & Burton, 2005; Dorans, 2004; Dorans & Zeller, 2004a, 2004b) responded to Freedle's (2003) work by attributing his findings to statistical artifacts, technical problems in the implementation of the methodology he used to identify DIF (Dorans & Kulick, 1983), and to the effects of student guessing. In their view, issues of improper scoring and score linking, in addition to the use of obsolete data, rendered

Freedle's (2003) findings insignificant. They also present analyses criticizing the validity and reliability of the scoring proposed by Freedle (2003) to mitigate the impact of these reported unfair results.

4. For a review of DIF methods see Millsap and Everson (1993). For other IRT approaches to DIF, see Hambleton et al. (1991), Holland and Wainer (1993), and Camilli and Shepard (1994).
5. Bridgeman and Burton (2005) make no reference to how to match students on "true ability" rather than on total score.
6. Which discourages guessing.
7. The phenomenon Freedle described was consistently present when comparing item performance between students from different ethnic groups whose preferred language was English. The correlation between item difficulty and DIF estimates became stronger when students whose preferred language was other than English were included in the analysis. For more recent evidence regarding the role of the first language in DIF results, see Sinharay, Dorans, and Liang (2009).
8. Uniform DIF occurs when the difference in probabilities of success is uniform for the two groups over all ability levels. In nonuniform-DIF situations the probabilities of success differ differently for people at higher ability levels and at lower ability levels (Hambleton et al., 1991).
9. It needs to be noted that no iterative screening for DIF items (Millsap & Everson, 1993) was conducted for the research reported here because it was assumed that SAT items would exhibit small DIF effect sizes thanks to the DIF screening process in place in ETS since the 1980s. Thus, the focus of this investigation is about the relationship between item difficulty and DIF estimates rather than on the effect sizes of the specific DIF estimates.
10. Some of these Rasch-type IRT models are the logistic Rasch model, the rating scale model, the partial credit model, and the ordered partition model.
11. They were divided by 2 because in this model each set of DIF parameters estimate the distance between the common item difficulty parameter and the item difficulty corresponding to that ethnic group. Paek (2002), on the other hand, estimated a model with no common item difficulty parameter. In his case the DIF estimates, therefore, measured the distance between the item difficulty parameters of both ethnic groups.
12. We also tried using Parscale (Scientific Software International, 2003) but it was not able to estimate the final deviance of models that would allow for DIF in all three parameters of the 3PL function simultaneously (parameters *a*, *b*, and *c*).
13. Unfortunately Thissen et al. (1988) did not work with the item discrimination parameter.
14. However, the hypothesized relationship between item difficulty and DIF estimates was not observed in the math test.
15. Item difficulty estimates from the reference group were used in an effort to follow closely Freedle's methodology. In Freedle's article, DIF estimates (standardization *p* index) were correlated with items' *p* values, or the proportion of examinees who responded to the item correctly. Since the reference group significantly outnumbers the focal group, the *p* value is mainly determined by the responses of the reference group. Results did not change when using item difficulty estimates from the focal group.

16. For the Math test of Form IZ, we observe a relative similar fit of the DIF and 1PL model. Additionally, the relationship between item difficulty and DIF estimates is weaker: The correlation between them is only 0.204 and it is not statistically significant ($p = .121$). See Table A2 in the appendix for results of the Math test by item type.
17. The analyses were also conducted on the Math test from forms DX, QI, IZ, and VD for both the White/African American and the White/Hispanic comparisons. The results show that the phenomenon described by Freedle was strong in the 1994 forms and it is weaker in more current Math tests. None of the results for the 1999 forms are statistically or practically significant ($p < .05$). See Table A3 in the appendix for details.
18. The two items that exhibit differences in the *c* parameter that classify for the largest category also show large DIF in the *b* parameter.

References

Adams, R. J., & Wilson, M. (1996). Formulating the Rasch model as a mixed coefficients multinomial logit. In G. Engelhard & M. Wilson (Eds.), *Objective measurement: Theory and practice* (Vol. 3, pp. 143-166). Norwood; NJ: Ablex.

Adams, R. J., Wilson, M. R., & Wang, M. L. (1997). The multidimensional random coefficients multinomial logit. *Applied Psychological Measurement*, *21*, 1-24.

Agresti, A., & Finlay, B. (1997). *Statistical methods for the social sciences*. Upper Saddle River, NJ: Prentice Hall.

Brennan, R. (Ed.). (2006). *Educational measurement*. Westport, CT: Praeger.

Bridgeman, B., & Burton, N. (2005, April). *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, Quebec, Canada.

Burton, E., & Burton, N. W. (1993). The effect of item screening on test scores and test characteristics. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 321-336). Hillsdale, NJ: Lawrence Erlbaum.

Camara, W., & Sathy, V. (2004). *College Board response to Harvard Educational Review article by Freedle*. Retrieved from http://professionals.collegeboard.com/profdownload/pdf/051425Harvard_050406.pdf

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items* (Vol. 4). London, England: Sage.

Carlton, S., & Harris, A. (1992, November). *Characteristics associated with differential item functioning on the Scholastic Aptitude Test: Gender and majority/minority group comparisons* (Report No. ETS-RR-92-64). Princeton, NJ: Educational Testing Service.

Clauser, B. E., & Mazor, K. M. (1998, Spring). *An NCME instructional module on "Using statistical procedures to identify differentially functioning test items*." Retrieved from http://www.ncme.org/pubs/items/Statistical.pdf

Cohen, A., & Bolt, D. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement*, *42*, 133-148.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum. (Original work published 1969)

De Boeck, P. (2010, October 26). Another look at bias in the SAT [Web log post]. Retrieved from http://www.hepg.org/blog/45

Dorans, N. (2004, Spring). Freedle's table 2: Fact or fiction. *Harvard Educational Review*, *74*, 62-73.

Dorans, N. (2010, September). *Unfair treatment vs. confirmation bias? Comments on Santelices and Wilson* (Report No. ETS-RR-10-20). Princeton, NJ: Educational Testing Service.

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* (Report No. RR-83-09). 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*, 355-368.

Dorans, N., & Lawrence, I. (1987). *The internal construct validity of the SAT* (Report No. RR-87-35). Princeton, NJ: Educational Testing Service.

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

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

Embretson, S., & Reise, S. (2000). *Item response theory for psychologists*. Hillsdale, NJ: Lawrence Erlbaum.

Freedle, R. (2003, Spring). Correcting the SAT's ethnic and social-class bias: A method for reestimating SAT scores. *Harvard Educational Review*, *73*, 1-43.

Freedle, R. (2010, Fall). On replicating ethnic test bias effects: The Santelices and Wilson study. *Harvard Educational Review*, *80*, 394-404.

Hambleton, R., Swaminathan, H., & Rogers, J. (1991). *Fundamentals of item response theory.* Thousand Oaks, CA: Sage.

Holland, P., & Thayer, D. T. (1986). *Differential item functioning and the Mantel-Haenszel procedure* (Report No. PSRTR-86-69). Princeton, NJ: Educational Testing Service.

Holland, P., & Thayer, D. T. (1988). Differential item performance and Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum. International Association for the Evaluation of Educational Achievement (IEA). IEA Civic Education Study, 1999: [United States] [Computer file]. ICPSR03892-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2004. doi:10.3886/ICPSR03892 (http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/03892/detail).

Holland, P., & Wainer, H. (1993). Concluding remarks and suggestions. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 419-438). Hillsdale, NJ: Lawrence Erlbaum.

International Association for the Evaluation of Educational Achievement (IEA). IEA Civic Education Study, 1999: [United States] [Computer file]. ICPSR03892-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2004. doi:10.3886/ICPSR03892 (http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/03892/detail).

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

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.

Longford, N., Holland, P., & Thayer, D. T. (1993). Stability of the MH D-DIF statistics across populations. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 171-196). Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Mathews, J. (2003, November). The bias question. *The Atlantic Monthly*, *292*(4), 130-140.

Millsap, R., & Everson, H. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, *17*, 297-334.

Mislevy, R. J., & Verhelst, N. (1990). Modeling item response when different subjects employ different solution strategies. *Psychometrika*, *55*, 195-215.

Moore, S. (1996). Estimating differential item functioning in the polytomous case with the random coefficients multinomial logits (RCML) model. In G. Englehard & M. Wilson (Eds.), *Objective Measurement III: Theory into practice* (pp. 219-239). Norwood, NJ: Ablex.

O'Neill, K., & McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 255-276). Hillsdale, NJ: Lawrence Erlbaum.

Paek, I. (2002). *Investigation of differential item functioning: Comparisons among approaches, and extension to a multidimensional context* (Unpublished doctoral dissertation). University of California, Berkeley.

Rao, C., & Sinharay, S. (Eds.). (2007). *Handbook of statistics: Vol. 26. Psychometrics*. Amsterdam, Netherlands: Elsevier.

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

Santelices, M. V., & Wilson, M. (2010a, Fall). Responding to claims of misrepresentation. *Harvard Educational Review*, *80*, 413-417.

Santelices, M. V., & Wilson, M. (2010b, Spring). Unfair treatment? The case of Freedle, the SAT and the standardization approach to differential item functioning. *Harvard Educational Review*, *80*, 106-134.

Scherbaum, C., & Goldstein, H. (2008). Examining the relationship between race-based differential item functioning and item difficulty. *Educational and Psychological Measurement*, *68*, 537-553.

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

Schmitt, A., & Dorans, N. (1988). *Differential item functioning for minority examinees on the SAT* (Report No. RR-88-32). Princeton, NJ: Educational Testing Service.

Scientific Software International. (2003). IRT from SSI: BILOG-MG, PARSCALE MULTILOG and TESTFACT [Computer software]. Lincolnwood, IL: Author.

Sinharay, S., Dorans, N., & Liang, L. (2009, March). *First language of examinees and its relationship to differential item functioning* (Report No. ETS RR-09-11). Princeton, NJ: Educational Testing Service.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 147-150). Hillsdale, NJ: Lawrence Erlbaum.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-114). Hillsdale, NJ: Lawrence Erlbaum.

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* (pp. 63-73). Princeton, NJ: Princeton University Press.

Wainer, H., & Skorupski, W. P. (2005). Was it ethnic and social-class bias or statistical artifact? Logical and empirical evidence against Freedle's method for reestimating SAT scores. *Chance*, *18*, 17-24.

Waller, W. I. (1989). Modeling guessing behavior. *Applied Psychological Measurement*, *13*, 233-243.

Wu, M., Adams, R. J., & Wilson, M. (1998). *ACER-ConQuest* [Computer software]. Hawthorn, Victoria, Australia: ACER Press.

Xie, Y. (2005). *Three studies of person by item interactions in international assessment of educational achievement* (Unpublished doctoral dissertation). University of California, Berkeley.

Yamamoto, K. (1987). *A model that combines IRT and latent class models* (Unpublished doctoral dissertation). University of Illinois, Champaign-Urbana.

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