UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

BRIDGING THE GAPS: AN EVALUATION OF THE GROUP ACTIVATED PROBABILITY OF SUCCESS MODEL OF STEREOTYPE THREAT

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

In partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By
MARIA COHENOUR
Norman, Oklahoma
2010

BRIDGING THE GAPS: AN EVALUATION OF THE GROUP ACTIVATED PROBABILITY OF SUCCESS MODEL OF STEREOTYPE THREAT

A DISSERTATION APPROVED FOR THE
DEPARTMENT OF PSYCHOLOGY

BY

_____
Dr. Robert Terry, Chair

_____
Dr. Joe Rodgers

_____
Dr. Ryan Brown

_____
Dr. Lara Mayeux

_____
Dr. Terry Pace

Table of Contents

List of Tables

List of Illustrations

Abstract

The authors developed the Group Activated Probability of Success (GAPS) model to replicate the relationship between the performances of minority group members that are impacted by stereotype threat and to examine the impact of stereotype threat from an item level perspective. To accomplish this, a traditional item response theory (IRT) model was modified to reflect the specific characteristics of the individual and the items that interact to influence the individual's estimated latent proficiency in the target domain. This resulted in non-uniform differential item functioning (DIF) in items of high difficulty for minority group members who are also of high proficiency in the target domain. The model was developed to simulate the effect of stereotype threat at the item level using item difficulty, individual latent proficiency, and group membership to alter the probability of success on a given item for individuals in the targeted group, controlling for proficiency in the target domain. Guided by stereotype threat research, the manipulation of these factors resulted in a model that successfully replicates the differences found in laboratory experiments and suggests a possible explanation for the lack of support for stereotype threat in applied research.

Bridging the GAPS: An Evaluation of the Group Activated Probability of Success Model of Stereotype Threat

Stereotype threat (ST) occurs when an individual's performance is negatively affected due to their group membership (e.g., race, gender) being made salient early in testing situations. This occurs when the individual is aware of a negative stereotype commonly associated with their group (e.g., women are not good at math), even if there is no truth to the stereotype. ST has been shown to attenuate test performance even under such minimal conditions as having the test-taker indicate their membership in a stereotyped group at the beginning of a test. For example, if we administer a math test to a female who is aware of the negative stereotype involving women and mathematical proficiency, ST research would predict that she would perform worse on the test if first asked to indicate her gender than if her gender had not been focused upon. ST can occur in a wide variety of domains; even members of groups normally stereotyped as adept may be susceptible under relevant circumstances (Aronson, Good, & Keough, 1999; Aronson, Quinn, Spencer, 1998; Spencer, Steele, & Quinn, 1999; Steele, 1997; Steele & Aronson, 1995).

One classic example (Shih, Pittinsky, & Ambaby 1999) manipulated the activation of two stereotypes: 1) women having difficulty with mathematics, and 2) Asians having superior mathematical proficiency. The researchers implicitly activated either gender identity or race identity in a group of Asian-American women prior to a mathematics test. Results indicated that the gender-activated group performed worse relevant to a control group, while the race-activated group performed better (a

phenomenon known as *stereotype lift*; Blanton, Buunk, Bibbsons, & Kuyper, 1999; Fein & Spencer, 1997).

In recent years, attention has been drawn to the value of ST theory in applied settings. Sackett, Hardison, and Cullen (2004) published a review that concluded that ST, when induced in the laboratory, could not account for differences between groups in the real world of high stakes testing. In a follow-up article, Cullen, Hardison, and Sackett (2004) evaluated four models of ST in several applied settings and found support for their previous contention that the effect of ST in real-world settings is minimal at best. In this paper, however, we contend that ST has been elusive in real-world settings due to a type of measurement bias that is undetectable using current standards of test evaluation.

If we accept the assumption that we should see stronger ST effects in applied settings than have been previously demonstrated (Aronson et al., 1998), the issues raised by Sackett et al. (2004) have the potential to devalue over ten years of laboratory research. However, one must consider the alternate explanation that the underlying mechanism by which ST influences test performance has yet to be identified. Nearly all studies investigating ST use the total test-score as the performance outcome measure; therefore, the possibility remains that a more sensitive analysis at a disaggregated level (i.e., the item-level) could potentially resolve the laboratory/applied setting debate. We propose that under traditional, applied testing conditions (e.g., the SAT), ST manifests itself at the item level in the form of non-uniform differential item functioning (DIF), which under current standards for test evaluation may not be detected. In sum, we believe that non-uniform differential item functioning is a valid explanation for the current finding that there is no evidence of ST in high stakes testing and applied ST research.

Two additional, common elements of the experimental literature may also work against the finding of ST in applied settings. First, it is typical to find that the impact is constrained only to items of significant difficulty (Spencer et al., 1999; Steele & Quinn, 1997). Second, the effect is greater for participants who demonstrate greater competence in the subject area being evaluated (Steele, 1997). The first component suggests that the difficulty of the item affects whether or not it will elicit differential performance. In other words, the vast majority of items (those that are not particularly difficult) will *not* elicit bias due to ST. The second component suggests that the proficiency of the individual is an essential factor in determining whether performance will be affected by ST. This, then, suggests that any one item may elicit differential performance for some individuals (those of high proficiency) but not others (those of low to moderate proficiency). Taken together, these findings are crucial because they describe the mechanisms by which a small set of items that exhibit non-uniform DIF are produced.

There are two major limitations that can be overcome by using an item level analysis (as opposed to the observed test score) to evaluate ST. First, when using only the observed test score, it is impossible to determine if there are any variations in the number and type of items that are influenced by ST. Using an item level analysis allows the researcher to take into account each item's contribution to the overall performance (see Figure 1). Second, the observed score does not present the opportunity to determine if there is an influence at the level of the individual's target ability. Implementing an item level analysis examines these potential variations and their role in predicting performance (see Figure 2). We propose that the observed test score is an insensitive measure for understanding the processes that influence test behavior as it relates to ST.

As a result, we argue that only item level models include the components necessary for clarifying ST effects in both experimental and applied settings.

The current study examined the impact of ST from an item level perspective by developing a model of the relationship between item level performances of minority group members that are likely to be influenced by ST. We developed the Group Activated Probability of Success (GAPS) model by modifying traditional item response theory (IRT) curves to reflect the specific characteristics of the individual as well as the items that interact to influence the individual's estimated proficiency while experiencing ST. IRT curves involve individuals' responses to items and, while there are many different types, the two-parameter logistic (2PL) model was used as a building block for the GAPS model presented in this paper. The 2PL IRT model is a mathematical function that relates the probability of an individual's item response to a characteristic of the individual (θ, proficiency level), and two characteristics of the item (α and β, discrimination and difficulty, respectively). The 2PL model assumes that the probability of an observable item response pattern can be linked to an estimate of an individual's position on an underlying latent variable (Hulin, Drasgow, & Parsons 1983; McDonald, 1999).

Fundamentally, the GAPS model simulates the influence of ST at the item level using item difficulty, individual proficiency, and group membership to alter the probability of success on a given item for individuals in the threatened group. We predict that a model manipulating these three factors will successfully replicate the ST effects found in laboratory experiments while suggesting a possible explanation for the lack of applied support for the theory.

While the GAPS model's assumptions are based on research supporting the general components of ST discussed previously, current research does not allow for the exploration of observed ST differences between the manipulation and control groups at an item level. The present project addresses this deficiency by incorporating elements of both ST and measurement theory into the proposed GAPS model. For our purposes, the most convenient way to introduce group differences into the mathematical equation at the item level is through the use of differential item functioning (DIF).

DIF was originally conceived as a method of detecting "bias" among test items. In the recent literature, DIF analyses have focused on the idea that distinct groups of examinees may react differently to test items, and this reaction may then affect performance at the item level (Holland & Wainer, 1993; Millsap & Meredith, 1992; Parshall & Miller, 1995; Scrams & McLeod, 2000; Williams, 1997). Systematic group differences in reaction to test items present a serious threat to the construct validity of tests, especially when such differences are not taken into account during the scoring process. The typical DIF analysis attempts to overcome this limitation by comparing two groups of examinees on an item suspected of DIF. Under these conditions, the performance of one group (the targeted group) is of primary interest while the performance of the other group (the non-targeted group) is used as a standard against which the performance of the targeted group can be compared. For example, in a DIF analysis of ST and math performance, the targeted group might be females and the non-targeted group males.

A critical feature of DIF is that it only evaluates *comparable members* of the targeted and non-targeted groups. These are determined by computing a conditional

probability on the construct of interest (e.g., mathematical proficiency) using either a known anchor set of DIF-free items or some proxy measure of the construct of interest. The DIF-free items used in the anchor set are typically items that have already been evaluated for, or are not suspected of, exhibiting DIF. These DIF-free items are then used to estimate the proficiency of members of both the targeted and non-targeted groups. Once these estimates are obtained, they serve to match the examinees from both groups on proficiency while simultaneously allowing the suspect items to be examined for DIF. Depending on both the quantity and mathematical properties of items displaying DIF, then, this analysis can be used to evaluate the possibility that the test as a whole (as opposed to specific items) may display systematic group differences.

DIF consists of two distinct analyses for determining how a single item or set of items is affecting the performance of the targeted group relative to the non-targeted group. The two possible DIF-related analyses are referred to as uniform DIF analysis and non-uniform DIF analysis (see Figure 3 and 4, respectively). An analysis of uniform DIF gives the researcher an estimate of a possible *constant difference* in item performance between groups while controlling for the proficiency of the examinee. This analysis assumes that DIF consistently favors one group regardless of proficiency level; it is analogous to the covariate-adjusted two-way interaction between the group and the item, using latent proficiency as the covariate. If constant DIF is operating within a given test, these item-specific estimates can be used to make statistical adjustments to test scores, which can reduce the impact of DIF on the total scores of the targeted group.

The second type of analysis is non-uniform or non-constant DIF. Non-uniform DIF occurs when two or more matched-proficiency groups show differential performance

on a given item, but the difference affects only individuals at certain levels of proficiency. The presence of non-uniform DIF is detected in an item when the item shows a non-constant performance difference between groups. Depending upon the overall difficulty of the item, the impact of non-uniform DIF on total test performance may be trivial because very few examinees will respond either correctly (to very difficult items) or incorrectly (to very easy items; Holland & Wainer, 1993). When non-uniform DIF impact occurs, it is analogous to a three-way interaction between group, item, and the covariate of latent proficiency.

Our goal in the current study was to create a DIF model to integrate the applied and laboratory-based research on ST with current measurement theory. It is our contention that non-uniform DIF is the underlying force driving the observed ST effects demonstrated in previous research. This explanation has the potential to impact an extensive body of literature, from laboratory studies demonstrating significant ST effects (Aronson et al. 1999; Aronson et al., 1998; Steele 1997) to the minimal results typically found in applied research (Cullen et al., 2004). The purpose of the current study is to demonstrate that the DIF model can account for ST effects within both the experimental and applied literature.

The current project is composed of two main components: 1) a meta-analysis of the experimental ST literature, specifically examining the impact of test difficulty and sample selection upon effect sizes, and 2) the development and evaluation of the GAPS model to reconcile both experimental and applied research findings.

Meta–Analysis of Stereotype Threat

Conducting a meta-analysis of ST research is a necessary preliminary step in understanding the true impact that the phenomenon has on test performance at both the item and test level. The effect sizes obtained from this analysis will be used to estimate parameters for the GAPS model.

Method

Retrieval of studies and inclusion criteria

To retrieve relevant studies for the meta-analysis, we employed inclusion criterion similar to that previously used by Walton and Cohen (2003). We first conducted a June 2007 search of the PsycINFO database using the words "stereotype threat," "testing," "gender," and "race". We then solicited additional studies by contacting experts in the field for additional studies that were not located using PsycINFO. The inclusion criteria we set required that studies evaluate the test performance of members of a negatively stereotyped group (e.g., females, African Americans, etc.). Participants had to be randomly assigned to one of at least three conditions: (1) an ST condition, (2) a control condition, or (3) a no ST condition (i.e., ST reduction condition). These conditions also had to accomplish an established manipulation goal. For the ST condition the stereotype in question had to be manipulated by the experimenter with the intention of increasing the level of threat experienced by the participant. The control condition was required to replicate the conditions that would be experienced by a participant in a real-world high stakes testing situation (to establish a baseline). Finally, in the ST reduction condition the stereotype in question had to be actively manipulated by the experimenter with the intention of decreasing the level of threat experienced by the participant. In sum, for studies to be considered, the performance measure used in the study had to be related to a

specific negative stereotype that was maximized in the ST condition, un-manipulated in the control condition, and reduced in the ST reduction condition.

In addition to the inclusion criteria used by Walton and Cohen (2003), we included studies with performance measures that ranged in difficulty from moderately difficult to very difficult. The inclusion of these additional studies was necessary to develop a more accurate description of how the relative difficulty of a performance measure impacts the measured effect size for various ST conditions.

Ultimately, 39 studies meeting the above criteria were included in this meta-analysis. For each study, we collected relevant descriptive statistics (e.g., mean, standard deviation, etc.) and calculated the size of the ST effect (*d*). Each included study had six possible effect sizes to be calculated, resulting in 234 possible effect size estimates obtained from this sample.[1] Each effect size came from a cell of the 3 (level of threat) x 2 (targeted versus non-targeted group) factorial design of our meta-analysis.

We established a coding system to test the following claims: 1) study design (as represented by the types and number of conditions included in an individual study) impacts the size of the ST effect and 2) the type of negative stereotype being tested (Males vs. Females, Whites vs. Blacks) does not impact the size of the ST effect (Lipsey & Wilson, 2001).

Coding stereotype threat studies

[1] Although there are 234 possible effect size estimates, the actual number of calculated effect sizes will be considerably smaller due to inconsistencies in the stereotype threat methodology that result in differences in types of participants and conditions used for the studies.

The studies included in our meta-analysis share three features: 1) they have one primary manipulated variable (the stereotype situation), 2) they have a pre-existing variable represented by the stereotype being tested (the impact group), and 3) a second pre-existing variable that is the group to which an individual participant identifies (either the targeted group or the non-targeted group). For example, if a researcher were interested in ST involving math ability (the stereotype situation) among females, then gender would be the impact group, females would be the targeted group, and males would be the non-targeted group. In other words, the impact group is always race, gender, or both, but the targeted group could be, for example, males or females, Blacks or Whites.

The ST situation has three possible levels: 1) ST, 2) control, 3) and ST reduction. The ST condition occurs when information is provided about the test that increases the threat experienced by participants (e.g., "gender differences have been found on this test"). The control condition occurs when no information is given about the nature of the test. This condition is typically expected to simulate conditions in a real-world testing environment. Finally, the ST reduction condition occurs when information is provided about the test to reduce the threat experienced by the participant (i.e. "no gender differences have been found for this test"). Table 1 contains the study type number used for identification in this study and corresponding design.

The two pre–existing variables are the impact group (e.g., gender) and the impact target (e.g., females). The coding for these two variables is interdependent because before the impact target can be coded the impact group must be established. The two most commonly used impact groups are gender and race; while some studies employ a mixture

of these two impact groups, only studies with a single impact group were used in the current meta-analysis (see inclusion criteria discussed previously). For studies using gender as the impact group, no impact target code was necessary; the targeted group was always "female" and the non-targeted group was always "male". However, for studies using race as the impact group, the targeted group needed to be coded to determine which race was the focus of the study and which race was to be used as a reference for comparative scores. Table 2 contains the codes for target group coding in studies with race as the impact factor.

Results and Discussion

Overview of sample

The key components of each study included in this meta-analysis are presented in Table 3. The summary statistics and codes are presented in Tables 4 and 5. As mentioned above, for each of the 39 studies included, a total of six possible effect sizes could be calculated from each complete study. Out of 234 possible effect sizes, it was possible to calculate 89 (60 targeted group effect sizes and 29 non-targeted group effect sizes). Of the calculated effect sizes, 73 (53 targeted group effect sizes and 20 non-targeted group effect sizes) showed the predicted pattern of results; Figure 5 depicts the pattern of effect sizes. It is evident that the effect sizes form two distinct distributions: one of ST (consisting of negative effect sizes), and one of stereotype lift (consisting of positive effect sizes). We evaluated these two distributions separately.

The distribution of ST effect sizes (the three effect sizes associated with the targeted group) were found to be negatively skewed with a mean of -0.49 (SD = 0.12, skewness = -2.27, kurtosis = 8.20). The distribution of stereotype lift effect sizes (the

three effect sizes associated with the non-targeted group) were found to be positively skewed with a mean of 0.10 (SD = 0.10, skewness = 1.79, kurtosis = 3.63). Additionally, of the six effect sizes, the threat versus threat reduction conditions showed the greatest effect sizes with a mean -0.56.

Results from the current meta-analysis revealed that the majority of effect sizes (73 out of 89) were in the predicted direction; this finding supports previous research conducted by Walton and Cohen (2003). The distribution of meta-analysis effect sizes were slightly negatively skewed for ST and slightly positively skewed for stereotype lift. Further, the absolute value distributions for ST and stereotype lift are mirror images of one another, suggesting that the distributions are similar in shape and variance.

Targeted and Non-Targeted Group Effect Sizes

To better understand how the overall effect size of each study is distributed over the impact groups, the effect size for each study was evaluated by impact group. The average effect size for the targeted group was significantly different from zero (d = -0.49, SD = 2.08; t (58) = -5.78, $p < 0.001$).

The targeted group effects were significant across all three levels of analysis (i.e. ST, control, and reduction conditions; see Table 6). In addition, post hoc tests (REGWQ) revealed that the effect size for the Threat-Reduction condition was significantly lower than the Threat-Control and Control-Reduction condition effect sizes. There was no difference between these latter two conditions.

Interestingly, the current data support the conjecture that the effect of ST is additive (Aronson, Good, & Keough, 1999; Aronson, Quinn, & Spencer, 1998; Spencer, Steele, & Quinn, 1999; Steele, 1997; Steele & Aronson, 1995). The level of threat

experienced in the threat reduction condition relative to the control condition is less than that experienced in the threat reduction condition relative to the ST condition. Further, the level of threat experienced when moving from the threat reduction condition to the control condition should be equivalent to that experienced when moving from the control condition to the ST condition. Additionally, the overall amount of ST experienced should be additive, such that [threat reduction to control] plus [control to ST] should equal [threat reduction to ST]. The data in the current meta-analysis support this idea (see Table 6).

The non-targeted group effect sizes were examined to demonstrate that the effect size for each study also was affected by impact group. The average effect size for the non-targeted group was not significantly different from zero; $d = 0.10$, *ns*. In addition, the non-targeted group effects were not found to be significant across any of the three comparison levels at an individual basis (see Table 7). In other words, the effect size for the Threat-Control condition was not significantly different than the effect size for either the Control-Reduction or Threat-Reduction condition ($d = 0.01$, SD = 0.07, *ns*; $d = 0.15$, SD = 0.12, *ns*, respectfully). There also was no difference between these latter two conditions ($d = 0.12$, SD = 0.09, *ns*). Due to the fact that none of the effect size estimates for the non-targeted group were significantly different from zero, we focused exclusively on the targeted group for the remainder of the analysis.

These results differ from those reported in Walton and Cohen's (2003) meta-analysis, which found an overall effect size difference for the non-target group. It is important to note, however, that their meta-analysis used stereotype lift as their primary variable of interest. Because of this difference in focus, the number of studies containing

non-targeted groups was considerably smaller in the current study. Therefore, it is not surprising that different results were obtained.

Overall tests of homogeneity of effect sizes and statistical significance

We used the total heterogeneity statistic to examine the homogeneity of our effect sizes (Wang & Bushman, 1999). The test revealed non-significant results (Q_T = 201.47, *ns*), which indicates that the effect size estimates used in this study have a random-effects variance that is approximately zero. This is a good indication that the effect sizes all come from a population of studies that are similar to one another.

The moderating role of study design and stereotype

As discussed earlier, the moderating role of two variables (study design and impact group) was assessed. As predicted, the impact group was not a significant moderator of the ST effect. The size of the effect did not differ significantly between the two groups (e.g., gender and race), $F(1, 59) = 1.59$, *ns*. Studies that used race as the targeted group yielded roughly the same effect size (d = -0.44, SD = 0.73, N = 11) as those using gender (d = -0.66, SD = 0.34, N = 49). This shows that the type of stereotype (race or gender) did not affect the size of the threat effect; this also confirms our previous finding that all studies included in this meta-analysis come from the same population of studies. However, there is a large difference in sample size between the two types of studies (N_{race} = 11; N_{gender} = 49). Though the means for these two groups are not significantly different, more studies that use race as the manipulated stereotype should be conducted and incorporated into this analysis before the role of study design as a moderating factor can be fully analyzed.

Test Difficulty and Selection Criteria

ST research often assumes that the effects of threat are limited to: 1) participants of high proficiency in the subject matter; and 2) tests that are comprised of very difficult items (Aronson et al., 1998; Spencer et al., 1999; Steele, 1997; Steele & Quinn, 1997). In order to better understand these assumptions, the current meta-analysis reviewed both factors. We divided participant proficiency into three categories: no selection (any participant not currently attending a college or university), college students (the participant must be attending a college or university but that college or university was not classified as an elite institution), and elite university students (the participant must be attending a college or university with at least an SATM $\geq$ 650 admissions criteria.). Participant selection was a significant predictor of effect size, $F(2, 39) = 4.93$, $p < 0.01$. Figure 6 illustrates the impact of subject selection among the three groups. The effect is most pronounced in the Elite University students, followed by college students, with the no selection group demonstrating the smallest overall effect size. These findings support the idea that the students that are of the greatest proficiency are the most impacted by ST.

Test difficulty was also divided into three categories: easy tests (average test-taker scoring above 60%), moderately difficult tests (average test-taker scored between 40% and 60%), and difficult tests (average test-taker scored less than 40%). As with participant selection, test difficulty also was found to be a significant predictor of effect size, $F(2, 39) = 7.15$, $p < 0.01$. Figure 7 displays the impact of test difficulty across all three groups of participants. Results demonstrate that test difficulty behaves as predicted; difficult tests have the greatest impact on ST; the impact then decreases as the tests become easier. Descriptive measures of the effect size of ST by both participant proficiency and test difficulty are shown in Tables 8 and 9, respectively. The data

confirm both major assumptions of the literature: the largest threat effects are achieved when difficult items are given to participants high in subject domain proficiency.

Modeling Stereotype Threat

According to the ST literature (Spencer et al., 1999; Steele, 1997; Steele & Quinn, 1997; Aronson, Quinn, & Steele, 1998) there are several individual differences that influence whether or not a person will experience ST. We attempted to model several of these differences at the item level. To accomplish this, we modeled the individual differences, the specific items that interact to influence the individual's item level performance, and ultimately their total score. The model below was developed to simulate ST at the item level using item difficulty, individual proficiency, and the impact of ST on the item and the individual to alter the probability of success on a given item for individuals in both the non-targeted and targeted groups. We created the Group Activated Probability of Success (GAPS) item-level model by modifying a traditional two-parameter IRT model to create a response probability gap between groups that would ultimately reflect the empirical gaps in the literature. The GAPS model is given as follows:

$$P(X_i \mid \theta) = [1 + \exp\{\alpha_i(\theta - \beta_i) - \delta\alpha'(\theta - \omega)\}]^{-1} \quad (1)$$

$P(X_i \mid \theta)$ is the conditional (on θ) probability of success on item i, and the parameters α, β, and θ have the traditional IRT parameter interpretations, such that α_i is the item discrimination or slope parameter, β_i is the item difficulty or threshold parameter, and θ is the latent proficiency parameter for a test-taker. In addition to the standard IRT parameters, the GAPS model contains four parameters that serve to model the ST effect. These parameters are α', which results in a change in item slope for the

targeted group whenever ST is activated and is assumed to be constant across items; ω, which is the person-level impact introduced by ST that inhibits performance according to pure proficiency and is assumed to have a constant effect across items; and finally two parameters, θ' and β', which along with group membership jointly activate the ST parameters through the indicator variable δ. The indicator variable activates the ST parameters whenever high-proficiency individuals of a targeted group encounter difficult items according to the following threshold model:

$$\delta = 1 \text{ if } \beta > \beta' \text{ and } \theta > \theta' \text{ and group=target, 0 otherwise} \quad (2)$$

Because the GAPS parameters are new in modeling item responses, we turn now to a discussion of the effects of including these parameters.

Delta (δ)

For the GAPS model to accurately replicate findings from the literature, a method must be employed for activating the ST parameters (α' and ω) when testing conditions match the selection criteria used by researchers. This is accomplished by creating an indicator variable that activates these parameters only under certain conditions. The conditions under which these parameters are activated involve two criteria: the level of item difficulty in the tests used to assess ST and differences in the distributions of test-taker proficiencies. Based on the literature, the items selected tend to be extremely difficult and the participants tend to be very proficient in the subject area of interest. In order to replicate these conditions, delta must only activate the ST parameters when a) the items are of an appropriate difficulty, b) the individual is of a certain proficiency level, and c) the individual is a member of the group for which the ST applies. Thus, two threshold parameters are needed to determine whether or not delta will activate the

'threat' portion of the model. These parameters are: a) the threshold for item selection (beta prime) and b) the threshold for proficiency selection (theta prime).

Beta prime (β')

According to ST research (Spencer, Steele, & Quinn, 1999; Steele, 1997), item difficulty seems to be a mediating factor in whether or not ST occurs. For example, an item that is extremely easy will be answered correctly by most individuals regardless of whether or not their performance is being artificially lowered due to ST. The purpose of including beta prime is to set the threshold for how difficult items must be before they activate an ST response. This parameter identifies a point on the proficiency scale at which an item is of sufficient difficulty to create differential performance between the targeted and the non-targeted group.

An item will be considered to be of sufficient difficulty to experience ST whenever the item difficulty exceeds a certain threshold (i.e., $\beta > \beta'$). The value of β' will be estimated by fitting the model to ST effect sizes obtained in the literature using the average difficulty of the tests (and thus average item difficulty) as a surrogate for β'.

Theta prime (θ')

As with item difficulty, research suggests (Spencer et al., 1999; Steele, 1997) that not all individuals within the targeted group are equally affected by ST. For example, an individual who is low in math proficiency will probably not identify with the subject domain and consequently will not be threatened by the stereotype, which provides a buffer against performance deficit. The purpose of including theta prime is to establish a lower-bound threshold for how proficient an individual must be before they are impacted by ST. It identifies the point on the proficiency scale at which an individual in the target

group is sufficiently proficient to experience ST, resulting in performance differences from an individual of equal proficiency in the non-targeted group.

Within the GAPS model an individual will be considered of sufficient proficiency to experience ST whenever their proficiency exceeds a certain threshold (i.e. $\theta > \theta'$). The value of θ' also will be estimated by fitting the model to ST effect sizes obtained in the literature using the listed descriptions of the sample participants with respect to test score distributions (e.g. SAT, ACT) as a surrogate for the θ' parameter.

When these two threshold parameters are introduced via delta, the model is restricted by allowing only certain combinations of items and individuals to be subjected to the ST effect. For example, when delta is activated (i.e. $\delta=1$), the probability of a correct response is calculated as:

$$P(X_i \mid \theta) = [1 + \exp\{\alpha_i(\theta - \beta_i) - \alpha'(\theta - \omega)\}]^{-1} \quad (3)$$

This allows for differences between groups in terms of their item response curves.

However, when delta is not activated (i.e. $\delta = 0$), the probability of a correct response reduces to the basic two parameter IRT item characteristic curve:

$$P(X_i \mid \theta) = [1 + \exp\{\alpha_i(\theta - \beta_i)\}]^{-1} \quad (4)$$

This allows only for identical item response curves for both groups. We now turn to an interpretation of the ST parameters.

Alpha prime (α')

Alpha prime represents the impact of ST activation on the item characteristics by adjusting the slope (α) of the item characteristic curve downward. If $\alpha'= 0$, no change in the slope of the item response curve is predicted for the targeted group; thus, there is no

effect of ST. As α' approaches α, the slope of the item response curve approaches zero for the target group, indicating no relationship between increasing proficiency and the probability of correctly answering the item; this effect is, of course, limited to the range of proficiency delineated by the value of theta prime. In psychometric terms, the larger the value of alpha prime, the more impact the ST portion of the model will have on the probability of answering an item correctly for highly proficient individuals, in the sense that the item loses its proficiency to discriminate between individuals of high enough proficiency. Again, α' was estimated by optimizing it to replicate the empirical effect sizes from the meta-analysis.

Omega (ω)

Whereas changing the item-slope parameter constitutes an item-level effect of ST upon a response probability, an additional parameter is needed to represent the effect of ST on an individual test-taker's proficiency score. The parameter in the GAPS model used to instantiate this effect is ω, which decreases the estimated proficiency level for a test-taker under the influence of ST. If ω = 0, then there is no decrease in proficiency due to ST; as ω increases, the effect of ST is to make the test-taker look "less proficient". As before, ω is assumed to have a constant effect given that ST activation has occurred, and its value was estimated to replicate the effect sizes from the meta-analysis.

Parameter Estimation

When researchers score tests dichotomously (correct answers receive a 1 and incorrect answers receive a 0), the sum of all the item scores is the observed score. In stereotype threat research this is the most commonly used outcome measure. If an individual participant were to take an exam a large number of times their observed scores

would cluster around one value; this value is known as the true score (Baker, 2001; McDonald, 1999). The formula for a true score is given as:

$$TS_j = \sum_{i=1} P_i(\theta_j) \quad (5)$$

where: TS is the true score of examinees with ability level θ_j.

i denotes an item and Pi(θj) depends on the particular item characteristic curve employed.

Using this formula we are able to calculate the true score for any participant along the ability scale. If we were to calculate a true score for all possible values along the ability scale these true scores could be plotted to form the test characteristic curve (TCC). The test characteristic curve describes the relationship between a participants ability and their true score(Baker, 2001; McDonald, 1999).

For example, the true score for a four-item test is calculated (at an ability level of 1.0) below using the 2PL IRT model.

Item 1:

$$P_1(1.0) = 1/(1 + EXP(1.0(1.0 – (1.2)))) = 0.45 \quad (6)$$

Item 2:

$$P_2(1.0) = 1/(1 + EXP(1.2(1.0 – (0.89)))) = 0.53 \quad (7)$$

Item 3:

$$P_3(1.0) = 1/(1 + EXP(1.4(1.0 – (0.5)))) = 0.67 \quad (8)$$

Item 4:

$$P_4(1.0) = 1/(1 + EXP(1.36(1.0 – (1.5)))) = 0.34 \quad (9)$$

The TCC for a participant with an ability of $\theta = 1.0$, can then be calculated by summing the probabilities of the four items:

$$TS = 0.45 + 0.53 + 0.67 + 0.34 = 1.98 \quad (10)$$

So for participants with a latent proficiency of 1.0 the true score would be 1.98. In item response theory the TCC provides a method for transforming ability scores into true scores. This is especially important in situations where the researcher may not be able to interpret ability scores. Because the true score is equivalent to the expected value of a given participants test score, the true score can be used as a proxy for individual test scores (Baker, 2001; McDonald, 1999).

We conducted the meta-analysis to confirm the assumptions that underlie and justify the GAPS model parameters, and to provide an idea of the general range or area in which those parameter values might lie. From the meta-analytic data and our knowledge of the behavior of parameters within a traditional IRT model, we created a discrete finite range for each of the four parameters needed to simulate ST. Based on the research and common parameter values used in standardized testing research, the following parameter ranges were established.

Alpha prime: range 0.1 to 0.5

This is a slope adjustment in the model, and therefore is directly related to the original IRT parameter that represents the slope alpha. Alpha typically has a parameter

value of 1.0 +/- 0.36 for items that would typically be found on a standardized test, making it unreasonable to let alpha prime vary beyond these maximum and the minimum values. Alpha prime represents a reduction in the slope or discrimination of an item; thus, for values of alpha prime higher than 0.5, the possibility exists that an item's slope could be reduced to the point where there is very little or no discrimination between individuals' performance on the item. This is inconsistent with research that shows sufficient variability among individuals experiencing ST.

Omega: range 0.1 to 1.5

This parameter represents the nuisance factor that causes a detriment in performance due to ST. The research suggests that this parameter should be part of the model, but there is little agreement on its cause or magnitude. Given this lack of knowledge, we searched a wide parameter space, and the little restriction that was placed on the range was based on the theta distribution to which omega is directly related. Typically, theta has a range of -4.0 to 4.0 with a mean of 0 and a standard deviation of 1.0. It is unlikely that students below the mean would be identified with any target domain; therefore the low end of the range was set above the mean of theta. Due to the fact that omega is a nuisance factor, we determined that it should not have more influence than 1.5 SD, leaving the top of the omega range set at 1.5.

Beta prime: range 0.5 to 1.5

We used this parameter to determine which items are susceptible to ST and which items function traditionally. The range was based on the results of the meta-analysis. One consistent feature of the studies reviewed was the selection of difficult test items to be used when trying to induce ST. The effects of the increasing difficulty of test items can

be seen in Figure 7. As a result, the beta prime range was based on the upper half of the theta range and was capped at 1.5 SD above the mean.

Theta prime: range 0.5 to 1.5

We used this parameter to determine which participants within the targeted group are susceptible to ST. The range was determined using the same principles as beta prime and was based on Figure 6.

Search for Optimal Parameters

The parameters for this model were estimated using PROC NLIN in SAS© version 9.1.3, and the BEST option was used to specify that a grid search be conducted to find the optimal values for each parameter within the ranges discussed previously ("Introduction to SAS," 2009). The minimum specification to fit a nonlinear regression with PROC NLIN demands that the researcher specify the model and its parameters. All terms in the model not defined as parameters should be found in the dataset processed by PROC NLIN.

There are two types of undefined parameters in PROC NLIN. The first consists of the effect sizes collected from the meta-analysis, which were used as the closest empirical version of a dependent variable for the model below. The second type consists of the three traditional IRT parameters, which include 20 item slopes and thresholds and 50 participant abilities (25 stereotyped and 25 control participants) that were generated to replicate the types of items typically used in ST research.

Item Generation

To select items for the traditional IRT parameters, two parameters were generated: 20 item thresholds (β) and 20 item slopes (α). Item difficulty was generated

based on the meta-analysis. In order to activate the GAPS model, item difficulty had to be $1.2 \leq \beta \leq 4.0$. In order to determine the discrimination of the items we used the parameter range suggested by Shealy and Stout (1993). They suggested that an item discrimination parameter should have a range of 1.0 +/- 0.35; therefore we generated this parameter uniformly with this range. In all, 20 items were generated (for each study) with a difficulty and discrimination within the ranges mentioned above.[2]

Participant Generation

To generate participants for the traditional IRT parameter, the participants' proficiency distribution was set to be a standard random normal distribution with a range of $1.0 \leq \theta \leq 4.0$. Using this range of possible theta values, 25 participants were selected for each condition.

In order to use the empirical data from the meta-analysis, which is in true score form, it was necessary to use the TCC methodology to transform the GAPS model into a TCC curve that could be used in the PROC NLIN procedure to estimate the parameter values. For simplicity, we have abbreviated the GAPS model and the standard 2PL IRT model as follows.

$$P_1(\theta) = P(X_i \mid \theta) = [1 + \exp\{\alpha_i(\theta - \beta_i) - \delta\alpha'(\theta - \omega)\}]^{-1} \quad (11)$$

$$P_0(\theta) = P(X_i \mid \theta) = [1 + \exp\{\alpha_i(\theta - \beta_i)\}]^{-1} \quad (12)$$

The TTC for the GAPS model for an individual participant (*i*) then becomes

$$T(\theta) = \sum_{i=1}^{k} P_1(\theta) \quad (13)$$

and the TTC for the standard 2PL IRT model for an individual participant (*i*) becomes

k

[2] We used 20 items to replicate the typical test length of studies included in the meta-analysis.

$$NT(\theta) = \sum_{i=1} P_0(\theta) \quad (14)$$

However, simply converting the two models using the TCC method is not sufficient. Because the data from the meta-analysis must be in the form of effect sizes in order for the model to correctly predict the effect size, it is necessary that we use the two TCC equations above for an effect size equivalent. In this case, a delta TCC, or ΔTCC, was formed by summing across all of the participant TCC curves and dividing by the total number of participants, then subtracting the standard 2PL IRT group from the GAPS group:

$$\Delta TCC = \frac{\sum_{p=1}^{n} T(\theta_p)}{N} \text{ -- } \frac{\sum_{p=1}^{n} NT(\theta_p)}{N} \quad (15)$$

Several sets of parameter estimates were produced from the PROC NLIN grid search, and the group of parameters that best estimated the ST effect sizes was selected for use in the GAPS model. The parameter values were estimated to be: $\alpha' = 0.4$, $\omega = 0.79$, $\beta' = 1.2$ and $\theta' = 1.0$; these values were found to have a Levenberg – Marquardt pseudo $R^2 = 0.72$ ("Introduction to SAS," 2009). For simplicity, the above parameter estimates were based on the mean effect sizes for all ST conditions. This was done as an alternative to estimating parameters for each ST effect size (i.e., ST vs. control, ST vs. stereotype reduction, and stereotype reduction vs. control) because not all of the effect sizes were significantly different from one another, as demonstrated in the meta-analysis. In addition, the stereotype lift effect sizes were assumed to have zero impact based on the

magnitude of the effect revealed by the meta-analysis; however it would be informative to estimate parameters for all six effect sizes individually in future research.

Sensitivity Analysis of GAPS Model

A sensitivity analysis is the study of how variation in the output (e.g., the TCC or true score) of a model can be apportioned quantitatively to different sources of variation in the input of the model. It was conducted to investigate robustness to variation in the parameters of the GAPS model. Using the model as a framework, we systematically investigated expected results under different parameter configurations (Salteli, Chan, & Scott, 2000). The GAPS parameters were the input factors of interest because the parameters associated with the 2PL IRT model have previously been evaluated by IRT researchers (McDonald, 1999). The ultimate goals of the analysis were (1) to evaluate the impact of varying α', β', θ', and ω over their full ranges, and (2) to evaluate how that variation impacts the TCC. In addition to varying the input variables, we needed to ensure that the parameters were robust under both conditions being replicated by the model (labortory versus applied research); hence, we also had to create sets of latent variable populations (described below).

We evaluated the above goals in the context of the observed score differences typically manifested on high-stakes tests. In other words, we addressed the question, "how much of the total test score difference is attributed to the ST effect, as opposed to actual differences between groups?" To answer this question, we generating a set of test items to simulate items found on a typical math version of the SAT.

In order to evaluate the model and compute the distribution of the true score differences between the groups, we approached the sensitivity analysis from two

perspectives -- applied settings and ST research. The applied testing situation was the primary variable of interest in the current study and was based on the SAT college entrance exam (Education Testing Service, 2009). The SAT is commonly used as a tool for assessing participant proficiency in ST research (Aronson et. al., 1998; Spencer et. al., 1999; Steele, 1997). It consists of a verbal reasoning section, a mathematics section, and a writing section. For the purposes of the current study we focused exclusively on the mathematics section. This component of the SAT served as the primary template for creating a realistic exam that could be used to evaluating the GAPS model.

Method

In order to evaluate the parameters of the GAPS model under the conditions that exist in both laboratory and applied research, we used two latent proficiency distributions when conducting the sensitivity analysis. We first had to consider two sets of assumptions: 1) that the targeted and the non-targeted groups have equal proficiency distributions in laboratory research, and 2) that the targeted and the non-targeted groups have unequal proficiency distributions in applied research. The latter assumption suggests that the targeted group comes from a distribution with a mean that is lower, by some degree (assumed to be reflected in mean SAT score differences) than that of the non-targeted group in the proficiency domain of interest (see Figures 8 and 9 for a visual representation of the two sets of distributions).

In order to address both of these assumptions, we conducted two separate analyses. The first examined a no score gap scenario, whereas the second examined a score gap created by shifting the targeted group mean. The procedures used to generate the participants and test items for each assumption are as follows.

Assuming No Score Gap

Simulated Population

The targeted and non-targeted group theta values were generated from identical standard normal distributions with the following limits:

$$-4.0 \leq \theta \leq 4.0 \text{ by } 0.1.$$

Simulated Test

Fifty-four items were generated to simulate the SAT math test. Item difficulty and item discrimination parameters were generated for each item. Item difficulty (β) was generated from a stratified standard normal distribution[3], and item discrimination (α) was generated uniformly with a range of 1.0 +/- 0.35 (Shealy & Stout, 1993).

Assuming Observed SAT Score Gap

Simulated Population

Theta values for targeted and non-targeted groups were generated from unequal (mean-shifted) standard normal distributions. Specifically, we used the documented SAT© (The College Board, 2007) mean score differences to shift the distributions between the targeted and non-targeted group, which in turn reflects the true differences between the groups. The targeted group's theta values were generated from a standard normal distribution with the following limits:

[3] Ten items were generated with an item difficulty of $0 \leq \beta < 0.5$ and $-0.5 \leq \beta < 0$. Eight items were generated with an item difficulty of $0.5 \leq \beta < 1.0$ and $-1.0 \leq \beta < -0.5$. Five items were generated with an item difficulty of $1.0 \leq \beta < 1.5$ and $-1.5 \leq \beta < -1.0$. Two items were generated with an item difficulty of $1.5 \leq \beta < 2.0$, $\beta \geq 2.0$, $-2.0 \leq \beta < -1.5$, and $\beta < -2.0$.

$$-4.53 \leq \theta \leq 3.88 \text{ by } 0.1$$

and the non-targeted group's theta values were generated from a standard normal distribution with the following limits:

$$-4.0 \leq \theta \leq 4.0 \text{ by } 0.1$$

Simulated Test

To simulate the current SAT math test, 54 items were generated. Item difficulty and item discrimination parameters were generated for each item. Item difficulty (β) was generated from a stratified standard normal distribution and the item discrimination (α) was generated uniformly with a range of 1 +/- 0.35 (Shealy & Stout, 1993).

Parameter Variation

Parameters for the sensitivity analysis were varied exhaustively across all four distributions. The α' distribution ranged from 0.1 to 0.5 by increments of 0.1. The ω distribution ranged from 0.1 to 1.5 by increments of 0.1. The β' distribution ranged from 0.5 to 1.5 by increments of 0.1. The θ' distribution ranged from 0.5 to 1.5 by increments of 0.1. Using the TCC methodology in conjunction with the variations in parameter estimations allowed us to see how true score differences (ΔTCC) between groups under different assumptions would vary as the parameters varied.

Results and Discussion

When we assumed no differences in latent proficiency between the targeted and non-targeted groups, there was roughly a 1.5 true score point difference (30 point SAT scale score difference) between the two groups (see Figure 10). In contrast, when we assumed that the difference in means on the SAT reflected a true difference in latent

proficiency between the two groups (i.e., they come from proficiency distributions that are similar but whose means have been shifted), there was approximately a 5 true score point difference (100 point SAT scale score difference) between the groups (see Figure 11). These findings confirm the GAPS model's ability to replicate expected group differences based on the assumptions of the group proficiency distributions. In addition, there was little variation in the raw score difference between the two groups as the parameter values fluctuated under both assumed populations. These findings provide reasonable evidence that the GAPS model appears to be robust to variation in its parameters. Tables 10 through 17 contain the medians, quartiles, and minimum and maximum difference values as each of the input parameters vary for *assuming no score gap* and assuming *observed SAT score gap* groups that are associated with these tables.

The data collected from the sensitivity analysis of the model provide initial support for several of the assumptions discussed previously. First, they provide evidence that the parameters are robust to variation, which adds to the validity of the model. Second, they provide an estimate for the proportion of group differences that can be attributed to ST. Finally, they add support to the hypothesis that applied researchers have had little success in finding evidence of ST because it affects very few individuals and items in applied settings.

While the above analysis has provided a great deal of insight into ST research in applied settings, a deeper investigation into ST in the laboratory is still needed. To address this issue, a replication of this research was conducted to test whether or not the estimated parameters would mimic the results of the meta-analysis.

Stereotype Threat Research Replication

We tested the model's ability to replicate laboratory experiments by generating simulated data based on the populations sampled in traditional ST research and then comparing that data to the meta-analytic data.

Method

In order to replicate the results of the meta-analysis, 500 studies were simulated, each containing two groups and three conditions. The first group was a targeted group for which the model was activated in the threat condition (the control and threat reduction conditions used the traditional 2PL IRT model to calculate probabilities). The second group was a non-targeted group whose probabilities also were calculated using the traditional 2PL IRT model for all three conditions (see Table 18). Each of the simulated studies had a single, unique, 20-item exam[4] that was generated using a highly selected sample of items (items of high difficulty). In addition, participants were generated using a highly selected sample (participants with high proficiency) for each simulated study.

Item Generation

In order to select items, two parameters were generated: item difficulty (β) and item discrimination (α). Item difficulty was generated based on the meta-analysis. To activate the GAPS model, item difficulty had to exceed $1.2 \leq \beta \leq 4.0$. Item discrimination was determined using the parameter range of 1.0 +/- 0.35 (Shealy & Stout, 1993). In all, 20 items were generated for each study, each with a difficulty and discrimination within the ranges mentioned above.

Participant Generation

[4] We used 20 items to replicate the typical test length of studies included in the meta-analysis.

To accurately replicate ST research in the laboratory, we needed to generate participants that would activate the GAPS model based on the proficiency at which ST begins to impact an individual's performance. Based on the parameter values, the participants' proficiency distribution was set to be a standard random normal distribution with a range of $1.0 \leq \theta \leq 4.0$. Using this range of possible theta values, 25 participants were selected for each group in each condition, for a total of 250 participants that were generated for each replicated study (see Table 18 for a visual representation of the study design and model assignment for each of the 500 simulated studies).

Results and Discussion

Overview of sample

All of the 500 studies showed the predicted pattern of results (Md = -0.39, M = -0.39, SD = 0.10, skewness = 0.08, kurtosis = -0.27), which were consistent with empirical findings[5]. Figure 12 depicts the distributions of both the replicated and the empirical data for comparison. Based on these results, the GAPS model appears to accurately reflect the experimental findings under the conditions set in this replication.

General Discussion

Previous research in the area of ST has called into question its ability to account for differences between groups in the world of high stakes testing (Cullen et al., 2004). This criticism represented an important step in ST research, but was limited in its scope and failed to address other potential explanations for this finding. The current study was

[5] There were variations in skewness and kurtosis between the empirical findings and the replicated data due to extreme observations. However, when these observations were removed all statistical movements became consistent (see Table 19).

designed to bridge this gap by exploring several questions that are crucial to the field of ST research. First, is there an underlying mechanism in ST research that makes the phenomenon undetectable in applied settings? Second, can that mechanism be modeled and replicated? Finally, can the observed score differences between minority groups in high stakes testing situations be attributed in part or whole to this mechanism?

We were able to determine the robustness of the ST effect by conducting a much-needed meta-analysis. This analysis allowed us to determine the exact impact of this phenomenon on performance differences, as well as to examine the potential influence of other moderating variables. In addition, we identified the importance of accounting for test difficulty and participant selection in the process of determining the effect of ST. Finally, we revealed that the number of individuals and items impacted by ST is smaller than originally proposed.

The way in which the researcher defines his or her target population is important in determining the effect of ST. This is especially true in applied settings. Given the very small population of individuals that are influenced by ST, applied researchers must choose between evaluating the masses and searching for bias among the few. The extreme conditions under which ST occur, namely highly selected participants and very difficult tests, reduce the affected population considerably. These findings provide an explanation for the lack of support Cullen and colleagues (2004) found for ST effects in applied settings; namely that they used unselected participants and tests that were likely norm-referenced. Norm-referenced tests primarily contain questions designed to evaluate participants of average ability, not items that would typically be sensitive to ST situations. Our sensitivity analysis suggested that, even with the strongest possible

manipulation, ST would amount to a 1 SD difference between groups in an applied setting. Using the parameters that we determined to best replicate ST research, we would expect to find a much weaker effect -- only about 1/3 SD, which likely would not be significant. This lack of effect is even further exaggerated by the fact that only about 15% of applied ST samples would be at risk for ST.

Given these findings, we also modeled and replicated ST as it occurs in laboratory settings. The GAPS model replicates ST effects at the item level using four parameters: alpha prime (the impact of ST at the item level), omega (the nuisance variable that reduces participant proficiency as a result of ST), beta prime (item activation), and theta prime (participant proficiency activation). These parameters were found to be robust to variation and to accurately replicate laboratory findings.

One important facet of the current research is that ST is neither the sole cause nor absent from the score differences found between groups on standardized tests. The sensitivity analysis suggests that a proportion of these score differences can be attributed to ST in the form of DIF. Thus, group differences are not being detected through conventional means because only a small population is being affected and only a small number of items are susceptible.

While the research presented here represents an important first step, further studies need to be conducted to fully understand the nature of ST. First, a replication of the applied ST research needs to be conducted to determine if the GAPS model is capable of reproducing applied research as well as it can reproduce laboratory research. Second, estimation of parameters and a replication of stereotype lift in both laboratory and applied settings are needed to fully test the flexibility of the GAPS model. Third, once the

stereotype lift parameters have been estimated, a sensitivity analysis of those parameters should be conducted to determine their robustness. Finally, it would be advantageous to attempt to detect ST in a real high stakes testing data set using filters to determine the exact level of difference between the groups attributable to ST. This type of research would allow for a comparison between the estimated differences obtained from the sensitivity analysis and actual applied findings.

Lord (1980) suggested that biased items might not be cause for alarm or an indication that a test should be re-evaluated. Rather, they might be an indication that the test is not "strictly unidimensonal." In the case of the SAT and other standardized tests, it is possible that gender and race differences are not entirely a reflection of differences in proficiency. Instead, as Lord suggested, these differences may be due to the influence of some undetermined outside influence. The research conducted here provides one piece of the puzzle by demonstrating one potential variable that influences group differences at both the item and individual level (Holland & Wainer, 1993).

References

Ambady, N., Paik, S. K., Steele, J., Owen-Smith, A., & Mitchell, J. P. (2003). Deflecting negative self-relevant stereotype activation: The effects of individuation. *Journal of Experimental Social Psychology, 40*, 401 – 408.

Ambady, N., Smith, M., Kim, A., & Pittinsky, T. L. (2001). Stereotype susceptibility in children: Effects of identity activation and quantitative performance. *Psychological Science, 12*, 385 – 390.

Aronson, J., Fried, C. B., & Good, C. (2001). Reducing the effects of stereotype threat on African American college students by shaping theories of intelligence. *Journal of Experimental Social Psychology, 38*, 113 – 125.

Aronson, J., & Inzlicht, M. (2004). The ups and downs of attributional ambiguity. *Psychological Science, 15*(2), 829 – 836.

Aronson, J., Lustina, M. J., Good, C., & Keough, K. (1999). When white men can't do math: Necessary and sufficient factors in stereotype threat. *Journal of Experimental Social Psychology, 35*, 29 – 46.

Aronson, J., Quinn, D. M., & Spencer, S. J. (1998). Stereotype threat and the academic underperformance of minorities and women. *Academic Press*, 83 – 103.

Baker, F (2001). The basics of item response theory. College Park, MD: ERIC publications.

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology, 71*, 230- 244.

Blanton, H., Buunk, B. P., Gibbions, F. X., & Kuyper, H. (1999). When better-than-others compare upward: Choice comparison and comparative evaluation as independent predictors of academic performance. *Journal of Personality and Social Psychology, 83*, 817 – 827.

Brown, R. P., & Day, E. A. (2006). The difference isn't black and white: Stereotype threat and the race gap on Raven's advanced progressive matrices. *Journal of Applied Psychology, 91*, 979 - 985.

Brown, R. P., & Josephs, R. A. (1999). A burden of proof: Stereotype relevance and gender differences in math performance. *Journal of Personality and Social Psychology, 76*, 246 – 257.

Cadinu, M., Maass, A., Frigerio, S., Impagliazzo, L., & Latinotti, S. (2003). Stereotype threat: The effect of expectancy on performance. *European Journal of Social Psychology, 33*, 267 – 285.

Cadinu, M., Maass, A., Lombardo, M., & Frigerio, S. (2006). Stereotype threat: The moderating role of locus of control beliefs. *European Journal of Social Psychology, 36*, 183 – 197.

Cadinu, M., Maass, A., Rosabianca, A, & Kiesner, J. (2005). Why do women underperform under stereotype threat? Evidence for the role of negative thinking. *Psychological Science, 16*, 572 – 578.

Coe, R. (2002). It's the effect size, stupid. What effect size is and why it is important. Paper presented at the Annual Conference of the British Educational Research Association, University of Exeter, England, 12-14 September 2002.

Cohen, G. L., Garcia, J., Apfel, N., & Master, A. (2006). Reducing the racial achievement gap: A social-psychological intervention. *Science, 313,* 1307 – 1310.

Cullen, M. J., Hardison, C. M., & Sackett, P. R. (2004). Using SAT-grade and ability-job performance relationships to test predictions derived from stereotype threat theory. *Journal of Applied Psychology, 89,* 220 – 230.

Collins, W. C., Raju, N. S., & Edwards, J. E. (2000). Assessing differential functioning in a satisfaction scale. *Journal of Applied Psychology, 85,* 451 – 461.

College Board: Mean SAT score of college-bound seniors, 1967-2006. Retrieved from http://professionals.collegeboard.com/profdownload/cbs-2007-Table-2_Mean-SAT-Scores-of-College-Bound-Seniors-1972_2007.pdf

Davies, P. G., Spencer, S. J., Quinn, D. M., & Gerhardstein, R. (2002). Consuming images: How television commercials that elicit stereotype threat can restrain women academically and professionally. *Personality and Social Psychology Bulletin, 28,* 1615 – 1628.

Davies, P. G., Spencer, S. J., & Steele, C. M. (2005). Clearing the air: Identity safety moderates the effects of stereotype threat on women's leadership aspirations. *Journal of Personality and Social Psychology, 88,* 276 – 287.

DeGoot, M. H., & Schervish, M. J. (2001). Probability and Statistics 3rd edition. Reading, MA: Addison Wesley, Publishers.

Dorans, N. J. & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the scholastic Kaptitude test. *Journal of Educational Measurement, 23,* 355 – 368.

Frantz, C. M., Cuddy, A. J. C., Burnett, M., Ray, H., & Hart, A. (2004). A threat in the

computer: The race implicit association test as a stereotype threat experience. *Personality and Social Psychology Bulletin, 30*, 1611 – 1624.

Fein, S., & Spencer, S. J. (1997). Prejudice as self-image maintenance: Affirming the self through derogating others. *Journal of Personality and Social Psychology, 73*, 31 – 44.

Feinstein, Z. S. (1995). Effects of differing item parameters on closed-interval DIF statistics. *Applied Psychological Measurement, 19*, 131 – 142.

Gonzales, P. M., Blanton, H., & Williams, K. J. (2002). The effects of stereotype threat and double-minority status on the test performance of Latino women. *Personality and Social Psychology Bulletin, 28*, 659 – 670.

Good, C., Aronson, J., & Inzlicht, M. (2003). Improving adolescents' standardized test performance: An intervention to reduce the effects of stereotype threat. *Applied Developmental Psychology, 24*, 645 – 662.

Holland, P. W. & Wainer, H. (Eds.). (1993). Differential Item Functioning. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Hulin, C., Drasgow, F., & Parsons, C. (1983), Item Response Theory: Application to Psychological Measurement. Homewood, IL: Dorsey Professional Series.

Hunter, J. E., & Schmidt, F. L. (2000). The delimma of group differences: Racial and gender bias in ability and achievement tests: Resolving the apparent paradox. *Psychology, Public Policy, and Law, 6*, 151 – 158.

Introduction to SAS. UCLA: Academic Technology Services, Statistical Consulting Group. Retrieved from http://www.ats.ucla.edu/stat/sas/notesz/

Inzlicht, M., & Ben – Zeev, T. (2000). A threatening intellectual environment: Why

females are susceptible to experiencing problem-solving deficits in the presence of males. *Psychological Science, 11,* 365 - 371.

Inzlicht, M., & Ben – Zeev, T. (2003). Do high-achieving female students underperform in private? The implications of threatening environments on intellectual processing. *Journal of Educational Psychology, 95,* 796 – 805.

Johns, M., Schmader, T., & Martens, A. (2005). Knowing is half the battle teaching stereotype threat as a means to improving women's math performance. *American psychological society, 16,* 175 – 179.

Kawakami, K., Dovidio, J. F., & Dijksterhuis, A. (2003). Effect of social category priming on personal attitudes. *Psychological Science, 14,* 315 – 319.

Keller, J. (2002). Blatant stereotype threat and women's math performance: Self handicapping as a strategic means to cope with obtrusive negative performance expectations. *Sex Roles, 47,* 193 – 198.

Kiefer, A. K., & Sekaquaptewa, D. (2007). Implicit stereotypes, gender identification, and math-related outcomes: A prospective study of female college students. *Psychological Science, 18,* 13 – 18.

Leyens, J. P., Desert, M., Croziet, J. C., & Darcis, C. (2000). Stereotype threat: Are lower status and history of stigmatization preconditions of stereotype threat? *Personality and Social Psychology Bulletin, 26,* 1189 – 1199.

Lipsey, M. W. & Wilson, D. B. (2001). Practical Meta-Analysis. Thousand Oaks, CA: Sage Publications.

McKay, P. F., Doverspike, D., Bowen-Hilton, D., & McKay, Q. D. (2003). The effects of demographic variables and stereotype threat on black/white differences in

cognitive ability test performance. *Journal of Business and Psychology, 18,* 1 – 14.

Maass, A., & Cadinu, M. (2003). Stereotype threat: When minority members underperform. *European Review of Social Psychology, 14,* 243 – 275.

Marx, D. M., Stapel, D. A., & Muller, D. (2005). We can do it: The interplay of construal orientation and social comparisons under threat. *Journal of Personality and Social Psychology, 88,* 432 – 446.

McDonald, M. (1999). Test Theory: A Unified Treatment. Mahwah, NJ: Laurence Erlbaum Associates, Publishers.

Millsap, R. E., & Meredith, W. (1992). Inferential conditions in the statistical detection of measurement bias. *Applied Psychological Measurement, 16,* 389 – 402.

Monahan, P. O., McHorney, C. A., Stump, T. E., & Perkins, A. J. (2007). Odds Ratio, Delta, ETS classification, and standardization measures of DIF magnitude for binary logistic regression. *Journal of Educational and Behavioral Statistics, 32,* 92 – 109.

O'Brien, L. T., & Crandall, C. S. (2003). Stereotype threat and arousal: Effects on women's math performance. *Personality and Social Psychology Bulletin, 29,* 782 – 789.

Osborne, J. W. (1997). Race and academic disidentification. *Journal of Educational Psychology, 89,* 728 – 735.

Osborne, J. W. (1999). Unraveling underachievement among African American boys from an identification with academics perspective. *The Journal of Negro Education, 68,* 555 – 565.

Osborne, J. W. (2001). Testing stereotype threat: Does anxiety explain race and sex differences in achievement? *Contemporary Educational Psychology, 26,* 291 – 310.

Osborne, J. W., & Walker, C. (2006). Stereotype threat, identification with academics, and withdrawal form school: Why the most successful students of colour might be most likely to withdraw. *Educational Psychology, 26,* 563 – 577.

Oswald, D. L., & Harvey, R. D. (2000). Hostile environments, stereotype threat, and math performance among undergraduate women. *Current Psychology, 19,* 338 – 356.

Parshall, C. G., & Miller, T. R. (1995). Exact versus asymptotic mantel-haenszel DIF statistics: A comparison of performance under small-sample conditions. *Journal of Educational Measurement, 32,* 302 – 316.

Quinn, D. M., & Spencer, S. J. (2001). The interference of stereotype threat with women's generation of mathematical problem-solving strategies. *Journal of Social Issues, 57,* 55 – 71.

Rosenthal, H. E. S., & Crisp, R. J. (2006). Reducing stereotype threat by blurring intergroup boundaries. *Personality and Social Psychology Bulletin, 32,* 501 – 511.

Roznowski, M., & Reith, J. (1999). Examining the measurement quality of tests containing differentially functioning items: Do biased items result in poor measurement. *Educational and Psychological Measurement, 59,* 248 – 269.

Sackett, P. R., Hardison, C. M., & Cullen, M. J. (2004). On interpreting stereotype threat as accounting for African American – white differences on cognitive tests.

American Psychologist, 59, 7 – 13.

Schmader, T. (2002). Gender identification moderates stereotype threat effects on women's math performance. *Journal of Experimental Social Psychology, 38*, 194 – 201.

Schmader, T., & Johns, M. (2003). Converging evidence that stereotype threat reduces working memory capacity. *Journal of Personality and Social Psychology, 85*, 440 – 452.

Schimel, J., Arndt, J., Banko, K. M., & Cook, A. (2004). Not all self-affirmations were created equal: The cognitive and social benefits of affirming the intrinsic (vs. extrinsic) self. *Social Cognition, 22*, 75 – 99.

Scrams, D. L., & McLeod, L. D. (2000). An expected response function approach to graphical differential item functioning. *Journal of Educational Measurement, 37*, 263 – 280.

Sekaquaptewa, D., & Thompson, M. (2003). Solo status, stereotype threat, and performance expectancies: Their effects on women's performance. *Journal of Experimental Social Psychology, 39*, 68 – 74.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika, 58*, 159 – 194.

Shih, M., Ambady, N., Richeson, J. A., Fujita, K., & Gray, H. M. (2002). Stereotype performance boosts: The impact of self-relevance and the manner of stereotype activation. *Journal of Personality and Social Psychology, 83*, 638 – 647.

Shih, M., Pittinsky, T. L., & Ambady, N. (1999). Stereotype susceptibility: Identity

salience and shifts in quantitative performance. *Psychological Science, 10,* 80 – 83.

Spencer, S. J., Fein, S., Wolfe, C. T., Fong, C., & Dunn, M. A. (1998). Automatic activation of stereotypes: The role of self-image threat. *Personality and Social Psychology Bulletin, 24,* 1139 – 1152.

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology, 35,* 4 – 28.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2004). Examining the effects of differential item (functioning and differential) test functioning on selection decisions: When are statistically significant effects practically important? *Journal of Applied Psychology, 89,* 497 – 508.

Steele, C. M. (1997). A threat in the air how stereotypes shape intellectual identity and performance. *American Psychologist, 52,* 613 – 629.

Steele, C. M. (1999). Thin ice "stereotype threat" and black college students. *The Atlantic Monthly, 284,* 44 – 54.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69,* 797 – 811.

Steele, C. M., Spencer, S. J., & Aronson, J. (2002). Contending with group image: The psychology of stereotype and social identity threat. *Advances in Experimental Social Psychology, 34,* 379 – 439.

Swanson, D. B., Clauser, B. E., Case, S. M., Nungester, R. J., & Featherman, C. (2002). Analysis of differential item functioning (DIF) using hierarchical logistic

regression models. *Journal of Educational and Behavioral Statistics, 27,* 53 – 75.

Tatsuoka, K. K., Linn, R. L., Tatsuoka, M. M., & Yamamoto, K. (1988). Differential item functioning resulting from the use of different solution strategies. *Journal of Educational Measurement, 25,* 301 – 319.

Terry, R. A. (1989). Contingency table approaches to the identification of differential item functioning. *Dissertation Abstracts International, 51,* 3181 – 3280.

Voelkl, K. E. (1996). Measuring students' identification with school. *Educational and Psychological Measurement, 56,* 760 – 770.

Walsh, M., Hickey, C., & Duffy, J. (1999). Influence of item content and stereotype situation on gender differences in mathematical problem solving. *Sex Roles, 41,* 219 – 240.

Walton, G. M., & Cohen, G. L. (2003). Stereotype lift. *Journal of Experimental Social Psychology, 39,* 456 – 467.

Wang, M. C., & Bushman, B. J. (1999). Integrating results through meta-analytic review using SAS software. NC: SAS Institute, Inc.

Wang, N., & Lane, S. (1996). Detection of gender-related differential item functioning in a mathematics performance assessment. *Applied Measurement in Education, 9,* 175 – 199.

Wheeler, S. C., Jarvis, W. B., & Petty, R. E. (2001). Think unto others: The self destructive impact of negative racial stereotypes. *Journal of Experimental Social Psychology, 37,* 173 – 180.

Wicherts, J. M. (2005). Stereotype threat research and the assumptions underlying analysis of covariance. *American Psychologist, 60,* 267 – 269.

Wicherts, J. M., Dolan, C. V., & Hessen, D. J. (2005). Stereotype threat and group differences in test performance: A question of measurement invariance. *Journal of Personality and Social Psychology, 89,* 696 – 716.

Williams, V. (1997). The "unbiased" anchor: Bridging the gap between DIF and item bias. *Applied Measurement in Education, 10*, 253 – 267.

Wout, D., Danso, H., Jackson, J., Spencer, S., & Leland, J. (2008). The many faces of threat: Is stereotype threat a group-threat of a self-threat? *Journal of Experimental Social Psychology, 44*, 792 – 799.

Zhang, Y., Dorans, N. J., & Matthews-Lopez, J. L. (2005). Using DIF dissection method to assess effects of item deletion. The College Board. New York: New York.

Table 1

Coding Scheme for Type of Study Design

Study Type Number	Conditions Represented
01	1 Impact Level/ Threat and Reduction
02	1 Impact Level/ Threat and Control
03	1 Impact Level/ Control and Reduction
04	1 Impact Level/ Threat, Control and Reduction
05	2 Impact Levels/ Threat and Reduction
06	2 Impact Levels/ Threat and Control
07	2 Impact Levels/ Control and Reduction
08	2 Impact Levels/ Threat, Control and Reduction

Table 2

Coding Scheme for Impact Level in Studies with Race as the Impact Group

Race Code	Race by level
01	Targeted: Black; Non-targeted: White
02	Targeted: White; Non-targeted: Asian
03	Targeted: Hispanic; Non-targeted: White
04	White Only
05	Black Only
06	Hispanic Only
07	Asian Only

Table 3

Key Components of Each Stereotype Threat Study Included

Study	Targeted Group	Stereotype	Dependant Measure
Ambady, Paik, Steele, Owen-Smith, & Mitchell (2003), Study 1	Female Students	Gender	Canadian Math Competition
Ambady, Paik, Steele, Owen-Smith, & Mitchell (2003), Study 2	Female Students	Gender	Canadian Math Competition
Aronson & Inzlicht (2004)	Black Students	Race	GRE Verbal
Brown & Day (2007)	Black Students	Race	Raven's Advanced Progressive Matrices
Cadinu, Maass, Lombardo, & Frigerio (2006)	Female Students	Gender	Logic Test
Cadinu, Maass, Rosabianca, & Kiesner (2005)	Female Students	Gender	GRE Math
Candinu, Maass, Frigerio, Impagliazzo, & Latinotti (2002), Study 1	Female Students	Gender	Math Test
Candinu, Maass, Frigerio, Impagliazzo, & Latinotti (2002), Study 2	Female Students	Gender	Math Test
Frantz, Cuddy, Burnet, Ray, & Hart (2004), Study 1	Black Students	Race	IAT
Frantz, Cuddy, Burnet, Ray, & Hart (2004), Study 2	Black Students	Race	IAT
Frantz, Cuddy, Burnet, Ray, & Hart (2004), Study 3	Black Students	Race	IAT
Good, Aronson, & Inzlicht (2003)	Female Students	Gender	Texas Assessment of Academic skills
Inzlicht & Ben - Zeev (2000), Study 1	Female Students	Gender	GRE Math
Inzlicht & Ben - Zeev (2000), Study 2	Female Students	Gender	GRE Math
Inzlicht & Ben - Zeev (2003)	Female Students	Gender	GRE Math
Keller (2002)	Female Students	Gender	GMAT
Marx, Stapel, & Muller (2005), Study 1	Female Students	Gender	GRE Math
Marx, Stapel, & Muller (2005), Study 2	Female Students	Gender	GRE Math

Marx, Stapel, & Muller (2005), Study 3	Female Students	Gender	GRE Math
McKay, Doverspike, Bowen - Hilton, & McKay (2003)	Black Students	Race	Raven's Advanced Progressive Matrices
Oswald & Harvey (2000), Study 1	Female Students	Gender	GRE Math
Oswald & Harvey (2000), Study 2	Female Students	Gender	GRE Math
Quinn & Spencer (2001)	Female Students	Gender	SAT Math
Rosenthal & Crisp (2006), Study 1	Female Students	Gender	Math Test
Rosenthal & Crisp (2006), Study 2	Female Students	Gender	Math Test
Rosenthal & Crisp (2006), Study 3	Female Students	Gender	Math Test
Schimel, Arndt, Banko, & Cook (2004), Study 1	Female Students	Gender	GRE Math
Schimel, Arndt, Banko, & Cook (2004), Study 2	Female Students	Gender	GRE Math
Seibt & Forster (2004)	Female Students	Gender	Word Selection Test
Sekaquaptewa & Thompson (2003)	Female Students	Gender	GRE Math
Shih, Ambady, Richeson, Fujita, & Gray (2002)	White Students	Race	SAT Math
Wicherts, Dolan, & David (2005), Study 1	Black Students	Race	Differential Aptitude Test
Wicherts, Dolan, & David (2005), Study 2	Female Students	Gender	SAT Math (Difficult Items)
Wicherts, Dolan, & David (2005), Study 2	Female Students	Gender	SAT Math (Easy Items)
Wicherts, Dolan, & David (2005), Study 3	Female Students	Gender	Arithmetic Ability Test
Wicherts, Dolan, & David (2005), Study 3	Female Students	Gender	Number Series Task
Wicherts, Dolan, & David (2005), Study 3	Female Students	Gender	Mathematics Word Problems
Wicherts, Dolan, & David (2005), Study 3	Female Students	Gender	Primary Mental Abilities Test
Wout, Danso, Jackson, Spencer, & Leland (2008), Study 1	Female Students	Gender	SAT Math

Table 4

Summary Statistics and Coding for Targeted Group Members

Study	Study Type	Stereotype	Race Code	Stereotype Manipulation	Threat/Target Mean	Threat/ Target SD	Threat/ Target N	Control/ Target Mean	Control/ Target SD	Control/ Target N	Reduction/ Target Mean	Reduction/ Target SD	Reduction/ Target N
Ambady, Paik, Steele, Owen-Smith, & Mitchell (2003), Study 1	6	race	1	stereotype	0.16	0.11	12	0.3	0.21	12	.	.	.
Ambady, Paik, Steele, Owen-Smith, & Mitchell (2003), Study 2	1	gender	.	stereotype	40.8	9.17	20	.	.	.	45	15.3	20
Aronson & Inzlicht (2004)	1	gender	.	stereotype	44.5	20.3	20				51.6	28	19
Brown & Day (2007)	6	gender	.	test	8.9	3.62	16	10.3	3.26	21	.	.	.
Cadinu, Maass, Lombardo, & Frigerio (2006)	4	gender	.	other	12.67	8.9	15	18.56	11.8	16	23.28	11.8	18
Cadinu, Maass, Rosabianca, & Kiesner (2005)	1	gender	.	stereotype	6.15	2.6	13	.	.	.	9.75	3.4	12
Candinu, Maass, Frigerio, Impagliazzo, & Latinotti (2002), Study 1	4	gender	.	other	5.23	5.86	13	13.46	8.7	13	16.58	7.37	12
Candinu, Maass, Frigerio, Impagliazzo, & Latinotti (2002), Study 2	1	race	5	other	4.2	1.35	25	.	.	.	4.92	1.12	25
Frantz, Cuddy, Burnet, Ray, & Hart (2004), Study 1	4	race	4	stereotype	81.24	154.14	33	119.16	176.3	31	174.62	221.35	34
Frantz, Cuddy, Burnet, Ray, & Hart (2004), Study 2	1	race	4	stereotype	161.14	195	24	.	.	.	298.54	291	22

Frantz, Cuddy, Burnet, Ray, & Hart (2004), Study 3	1	gender	.	other	1.4	1.54	15	.	.	.	2.85	1.26	15
Good, Aronson, & Inzlicht (2003)	4	gender	.	other	4.8	1.55	12	5.43	1.95	12	6.83	1.8	12
Inzlicht & Ben - Zeev (2000), Study 1	1	gender	.	stereotype	6.2	1.9	15	.	.	.	7.58	1.78	16
Inzlicht & Ben - Zeev (2000), Study 2	6	gender	.	stereotype	5	1.37	16	5.57	1.63	15	.	.	.
Inzlicht & Ben - Zeev (2003)	5	race	2	stereotype	6.98	2.61	30	.	.	.	6.25	2.12	30
Keller (2002)	5	gender	.	other	0.58	0.03	34	.	.	.	0.7	0.04	34
Marx, Stapel, & Muller (2005), Study 1	1	gender	.	other	0.55	0.05	34	.	.	.	0.7	0.05	34
Marx, Stapel, & Muller (2005), Study 2	9	race	1	other	15.09	6.98	45	.	.	.	.	.	.
Marx, Stapel, & Muller (2005), Study 3	1	gender	.	stereotype	20.05	6.38	19	.	.	.	24.07	4.38	15
McKay, Doverspike, Bowen - Hilton, & McKay (2003)	1	gender	.	stereotype	19.45	4.56	20	.	.	.	22.06	5.05	18
Oswald & Harvey (2000), Study 1	1	gender	.	test	15.47	2.07	18	.	.	.	13.56	2.8	18
Oswald & Harvey (2000), Study 2	4	gender	.	test	11.3	2.42	16	11.93	2.87	15	13.75	2.41	15
Quinn & Spencer (2001)	1	gender	.	test	11.91	1.58	25	.	.	.	12.07	2.13	26
Rosenthal & Crisp (2006), Study 1	1	gender	.	test	10.67	2.9	38	.	.	.	13.83	1.79	39
Rosenthal & Crisp (2006), Study 2	1	gender	.	test	10.53	2.95	16	.	.	.	13.59	1.97	16
Rosenthal & Crisp (2006), Study 3	5	gender	.	stereotype	13.88	2.75	40	.	.	.	15.47	2.38	40
Schimel, Arndt, Banko, & Cook (2004), Study 1	8	gender	.	test	2.61	1.05	26	2.69	1.13	26	3.67	1.13	18

Schimel, Arndt, Banko, & Cook (2004), Study 2	7	gender	.	stereotype	.	.	.	74	6.94	28	82.11	5.72	28
Seibt & Forster (2004)	8	race	1	stereotype	19.41	5.6	17	22.42	5.15	19	24.29	5.05	17
Sekaquaptewa & Thompson (2003)	1	gender	.	stereotype	8.21	2.23	25	.	.	.	9.56	2.25	25
Shih, Ambady, Richeson, Fujita, & Gray (2002)	1	gender	.	stereotype	3.93	1.83	30	.	.	.	4.87	1.56	30
Wicherts, Dolan, & David (2005), Study 1	5	gender	.	stereotype	4.64	1.94	16	.	.	.	7.05	1.42	17
Wicherts, Dolan, & David (2005), Study 2	5	race	1	test	4.67	2.52	73	.	.	.	4.88	2.47	65
Wicherts, Dolan, & David (2005), Study 2	5	gender	.	test	6.81	2.55	28	.	.	.	7.99	2.88	30
Wicherts, Dolan, & David (2005), Study 3	5	gender	.	test	8.18	3.98	28	.	.	.	6.37	3.91	30
Wicherts, Dolan, & David (2005), Study 3	8	gender	.	test	9.96	6.16	47	10.23	4.62	48	11.7	3.53	47
Wicherts, Dolan, & David (2005), Study 3	8	gender	.	test	5.62	2.35	47	7.6	2.86	48	7.11	2.66	47
Wicherts, Dolan, & David (2005), Study 3	8	gender	.	test	11.81	5.18	47	11.55	5.14	48	11.21	4.66	47
Wout, Danso, Jackson, Spencer, & Leland (2008), Study 1	8	gender	.	test	5.74	2.72	47	6.4	2.8	48	6.72	2.32	47

Table 5

Summary Statistics and Coding for Non-Targeted Group Members

Study	Study Type	Stereotype	Race Code	Stereotype Manipulation	Threat/ Non-Target Mean	Threat/ Non-Target SD	Threat/ Non-Target N	Control/ Non-Target Mean	Control/ Non-Target SD	Control/ Non-Target N	Reduction/ Non-Target Mean	Reduction/ Non-Target SD	Reduction/ Non-Target N
Ambady, Paik, Steele, Owen-Smith, & Mitchell (2003), Study 1	6	race	1	stereotype	0.37	0.23	22	0.37	0.23	22	.	.	.
Ambady, Paik, Steele, Owen-Smith, & Mitchell (2003), Study 2	1	gender	.	stereotype	.	.	.	.	.	.	.	.	.
Aronson & Inzlicht (2004)	1	gender	.	stereotype	.	.	.				.	.	.
Brown & Day (2007)	6	gender	.	test	12.4	2.63	16	11.6	3.58	22	.	.	.
Cadinu, Maass, Lombardo, & Frigerio (2006)	4	gender	.	other	.	.	.	.	.	.	.	.	.
Cadinu, Maass, Rosabianca, & Kiesner (2005)	1	gender	.	stereotype	.	.	.	.	.	.	.	.	.
Candinu, Maass, Frigerio, Impagliazzo, & Latinotti (2002), Study 1	4	gender	.	other	.	.	.	.	.	.	.	.	.
Candinu, Maass, Frigerio, Impagliazzo, & Latinotti (2002), Study 2	1	race	5	other	.	.	.	.	.	.	.	.	.
Frantz, Cuddy, Burnet, Ray, & Hart (2004), Study 1	4	race	4	stereotype	.	.	.	.	.	.	.	.	.
Frantz, Cuddy, Burnet, Ray, & Hart (2004), Study 2	1	race	4	stereotype	.	.	.	.	.	.	.	.	.

Frantz, Cuddy, Burnet, Ray, & Hart (2004), Study 3	1	gender	.	other	.	.	.	.	.	.	.	.	.
Good, Aronson, & Inzlicht (2003)	4	gender	.	other						.	.	.	.
Inzlicht & Ben - Zeev (2000), Study 1	1	gender	.	stereotype	.	.	.	.	.	.	.	.	.
Inzlicht & Ben - Zeev (2000), Study 2	6	gender	.	stereotype	5.13	1.73	15	4.07	1.07	14	.	.	.
Inzlicht & Ben - Zeev (2003)	5	race	2	stereotype	9.12	2.35	15	.	.	.	5.14	1.32	15
Keller (2002)	5	gender	.	other	0.67	0.04	34	.	.	.	0.66	0.04	12
Marx, Stapel, & Muller (2005), Study 1	1	gender	.	other	.	.	.	.	.	.	.	.	.
Marx, Stapel, & Muller (2005), Study 2	9	race	1	other	19.79	6.51	42	.	.	.	.	.	.
McKay, Doverspike, Bowen - Hilton, & McKay (2003)	1	gender	.	stereotype	.	.	.	.	.	.	.	.	.
Oswald & Harvey (2000), Study 1	1	gender	.	test	.	.	.	.	.	.	.	.	.
Oswald & Harvey (2000), Study 2	4	gender	.	test	.	.	.	.	.	.	.	.	.
Quinn & Spencer (2001)	1	gender	.	test	.	.	.	.	.	.	.	.	.
Rosenthal & Crisp (2006), Study 1	1	gender	.	test	.	.	.	.	.	.	.	.	.
Rosenthal & Crisp (2006), Study 3	5	gender	.	stereotype	14.54	3.24	38	.	.	.	14.54	3.18	39
Schimel, Arndt, Banko, & Cook (2004), Study 1	8	gender	.	test	3.64	1.51	22	4	1.26	24	2.83	1.26	12
Schimel, Arndt, Banko, & Cook (2004), Study 2	7	gender	.	stereotype	.	.	.	81.9	5.95	42	81.55	6.03	42
Seibt & Forster (2004)	8	race	1	stereotype	24.67	3.89	27	24.28	4.08	29	22.44	4.56	27

Sekaquaptewa & Thompson (2003)	1	gender	.	stereotype	.	.	.	.	.	.	.	.	.
Shih, Ambady, Richeson, Fujita, & Gray (2002)	1	gender	.	stereotype	.	.	.	.	.	.	.	.	.
Wicherts, Dolan, & David (2005), Study 1	5	gender	.	stereotype	8.17	2.1	17	.	.	.	6.03	1.95	17
Wicherts, Dolan, & David (2005), Study 2	5	race	1	test	5.49	2.31	78	.	.	.	5.35	2.54	79
Wicherts, Dolan, & David (2005), Study 2	5	gender	.	test	9.19	2.51	51	.	.	.	9.13	2.36	50
Wicherts, Dolan, & David (2005), Study 3	5	gender	.	test	7.8	3.93	51	.	.	.	7.5	4.34	50
Wicherts, Dolan, & David (2005), Study 3	8	gender	.	test	12.2	5.33	45	13.28	7.46	46	14.18	7.78	50
Wicherts, Dolan, & David (2005), Study 3	8	gender	.	test	9.22	3.33	45	8.52	3.74	46	8.56	4.36	50
Wicherts, Dolan, & David (2005), Study 3	8	gender	.	test	12.97	5.11	45	12.9	5.92	46	13.14	5.86	50
Wout, Danso, Jackson, Spencer, & Leland (2008), Study 1	8	gender	.	test	7.44	2.88	45	8.39	3.43	46	7.6	3.09	50

Table 6

Mean Effect Size, Standard Deviation and Sample Size for the Targeted Group

Condition	Mean	SD	N
T - C	-0.32*	0.06	14
C - R	-0.24*	0.09	12
T - R	-0.56*	0.15	34

* significant at the $p < 0.05$ level

Table 7

Mean Effect Size, Standard Deviation, and Sample Size for the Non-Targeted Group

Condition	Mean	SD	N
T - C	0.01	0.07	9
C - R	0.12	0.09	7
T - R	0.15	0.12	13

* significant at the $p < 0.05$ level

Table 8

Target Group Effect Sizes by Participant Selection Criteria

Selection Criteria	Average Effect Size	SD	N
No Selection	-0.32	0.11	4
College Students	-0.52	0.08	28
College Students SATM > 650	-1.21	0.22	10

Table 9

Target Group Effect Sizes by Test Difficulty

Test Difficulty	Average Effect Size	SD	N
< 40%	-1.16	0.2	12
40% - 60%	-0.61	0.05	12
> 60 %	-0.22	0.09	18

Table 10

Quartiles and Extreme Measures for the Raw Score Difference as Alpha Prime Varies Assuming Equal Groups

	Alpha Prime				
Statistic	0.1	0.2	0.3	0.4	0.5
median	0.03	0.06	0.09	0.11	0.12
q1	0.01	0.03	0.03	0.04	0.05
min	-0.46	-0.95	-1.46	-2.00	-2.56
max	0.19	0.35	0.51	0.64	0.77
q3	0.06	0.12	0.16	0.20	0.23

Table 11

Quartiles and Extreme Measures for the Raw Score Difference as Alpha Prime Varies Assuming Unequal Groups

	Alpha Prime				
Statistic	0.1	0.2	0.3	0.4	0.5
median	-1.87	-1.87	-1.87	-1.87	-1.87
q1	-2.65	-2.70	-2.70	-2.70	-2.70
min	-3.11	-3.55	-4.01	-4.50	-5.00
max	-0.17	-0.12	-0.08	-0.06	-0.04
q3	-1.10	-1.08	-1.08	-1.06	-1.02

Table 12

Quartiles and Extreme Measures for the Raw Score Difference as Omega Varies Assuming Equal Groups

	Omega															
Statistic	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2	1.3	1.4	1.5
median	0.06	0.06	0.06	0.06	0.07	0.07	0.07	0.07	0.07	0.08	0.08	0.08	0.08	0.08	0.08	0.08
q1	0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.04	0.04	0.04	0.04	0.04
min	-2.56	-2.30	-2.05	-1.82	-1.59	-1.37	-1.16	-0.97	-0.80	-0.64	-0.49	-0.37	-0.25	-0.16	-0.07	0.00
max	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.76	0.75	0.74
q3	0.13	0.14	0.14	0.14	0.14	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.16	0.16	0.16	0.15

Table 13

Quartiles and Extreme Measures for the Raw Score Difference as Omega Varies Assuming Unequal Groups

	Omega															
Statistic	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2	1.3	1.4	1.5
median	-1.86	-1.87	-1.87	-1.87	-1.87	-1.87	-1.87	-1.87	-1.87	-1.87	-1.87	-1.87	-1.88	-1.89	-1.90	-1.92
q1	-2.65	-2.65	-2.65	-2.65	-2.65	-2.65	-2.67	-2.70	-2.70	-2.70	-2.70	-2.70	-2.70	-2.70	-2.70	-2.70
min	-3.01	-3.01	-3.01	-3.01	-3.14	-3.30	-3.45	-3.61	-3.78	-3.95	-4.12	-4.29	-4.47	-4.64	-4.82	-5.01
max	-0.04	-0.04	-0.05	-0.05	-0.05	-0.05	-0.06	-0.06	-0.06	-0.06	-0.07	-0.07	-0.07	-0.07	-0.08	-0.08
q3	-1.06	-1.07	-1.08	-1.08	-1.08	-1.08	-1.08	-1.08	-1.08	-1.08	-1.08	-1.08	-1.08	-1.09	-1.10	-1.10

Table 14

Quartiles and Extreme Measures for the Raw Score Difference as Beta Prime Varies Assuming Equal Groups.

	Beta Prime										
Statistic	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2	1.3	1.4	1.5
median	0.19	0.15	0.15	0.14	0.11	0.08	0.08	0.05	0.03	0.03	0.02
q1	0.11	0.09	0.09	0.08	0.06	0.05	0.05	0.03	0.02	0.01	0.01
min	-2.56	-2.19	-2.19	-2.01	-1.66	-1.15	-1.15	-0.82	-0.51	-0.51	-0.37
max	0.77	0.62	0.62	0.56	0.45	0.30	0.30	0.20	0.12	0.12	0.08
q3	0.29	0.24	0.24	0.21	0.17	0.12	0.12	0.07	0.05	0.05	0.03

Table 15

Quartiles and Extreme Measures for the Raw Score Difference as Beta Prime Varies Assuming Unequal Groups

	Beta Prime										
Statistic	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2	1.3	1.4	1.5
median	-1.87	-1.87	-1.90	-1.87	-1.87	-1.87	-1.87	-1.87	-1.88	-1.88	-1.88
q1	-2.70	-2.70	-2.70	-2.70	-2.70	-2.69	-2.70	-2.67	-2.65	-2.65	-2.65
min	-5.01	-4.67	-4.67	-4.50	-4.18	-3.72	-3.72	-3.42	-3.15	-3.15	-3.02
max	-0.04	-0.08	-0.08	-0.10	-0.13	-0.15	-0.15	-0.20	-0.21	-0.21	-0.23
q3	-1.02	-1.04	-1.04	-1.07	-1.08	-1.08	-1.08	-1.10	-1.12	-1.12	-1.14

Table 16

Quartiles and Extreme Measures for the Raw Score Difference as Theta Prime Varies Assuming Equal Groups

	Theta Prime															
Statistic	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2	1.3	1.4	1.5
median	0.12	0.11	0.11	0.10	0.09	0.09	0.08	0.08	0.07	0.06	0.06	0.05	0.05	0.04	0.04	0.03
q1	0.06	0.05	0.05	0.05	0.04	0.04	0.04	0.03	0.03	0.03	0.02	0.02	0.02	0.01	0.01	0.01
min	0.00	-0.15	-0.30	-0.46	-0.62	-0.78	-0.95	-1.12	-1.29	-1.46	-1.64	-1.82	-2.00	-2.19	-2.37	-2.56
max	0.77	0.72	0.67	0.63	0.59	0.55	0.51	0.48	0.44	0.41	0.38	0.36	0.33	0.31	0.29	0.27
q3	0.23	0.22	0.21	0.20	0.19	0.18	0.17	0.16	0.15	0.14	0.13	0.12	0.11	0.10	0.09	0.08

Table 17

Quartiles and Extreme Measures for the Raw Score Difference as Theta Prime Varies Assuming Unequal Groups

	Theta Prime															
Statistic	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2	1.3	1.4	1.5
median	-1.87	-1.87	-1.87	-1.87	-1.87	-1.87	-1.87	-1.87	-1.87	-1.87	-1.87	-1.87	-1.88	-1.88	-1.88	-1.88
q1	-2.71	-2.71	-2.71	-2.70	-2.70	-2.66	-2.65	-2.65	-2.65	-2.65	-2.65	-2.65	-2.65	-2.65	-2.65	-2.65
min	-5.01	-4.68	-4.68	-4.35	-4.04	-3.73	-3.43	-3.15	-3.01	-3.01	-3.01	-3.01	-3.01	-3.01	-3.01	-3.01
max	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04
q3	-1.08	-1.08	-1.08	-1.08	-1.08	-1.08	-1.08	-1.08	-1.08	-1.08	-1.08	-1.08	-1.08	-1.08	-1.08	-1.08

Table 18

Simulated Study Design for Stereotype Threat Replication

Group Identification	Stereotype Threat	Control	Stereotype Reduction
Targeted	N = 25 GAPS	N = 25 2 PL IRT	N = 25 2PL IRT
Non-Targeted	N = 25 2PL IRT	N = 25 2PL IRT	N = 25 2PL IRT

Table 19

Descriptive Statistics for Empirical, Adjusted, and Replicated Data

Descriptive Statistics	Empirical Data	Adjusted Data	Replicated Data
Mean	-0.49	-0.39	-0.39
Median	-0.36	-0.36	-0.39
SD	0.12	0.08	0.1
N	60	57	500
Skewness	-2.27	-0.13	0.08
Kurtosis	8.19	-0.14	-0.27

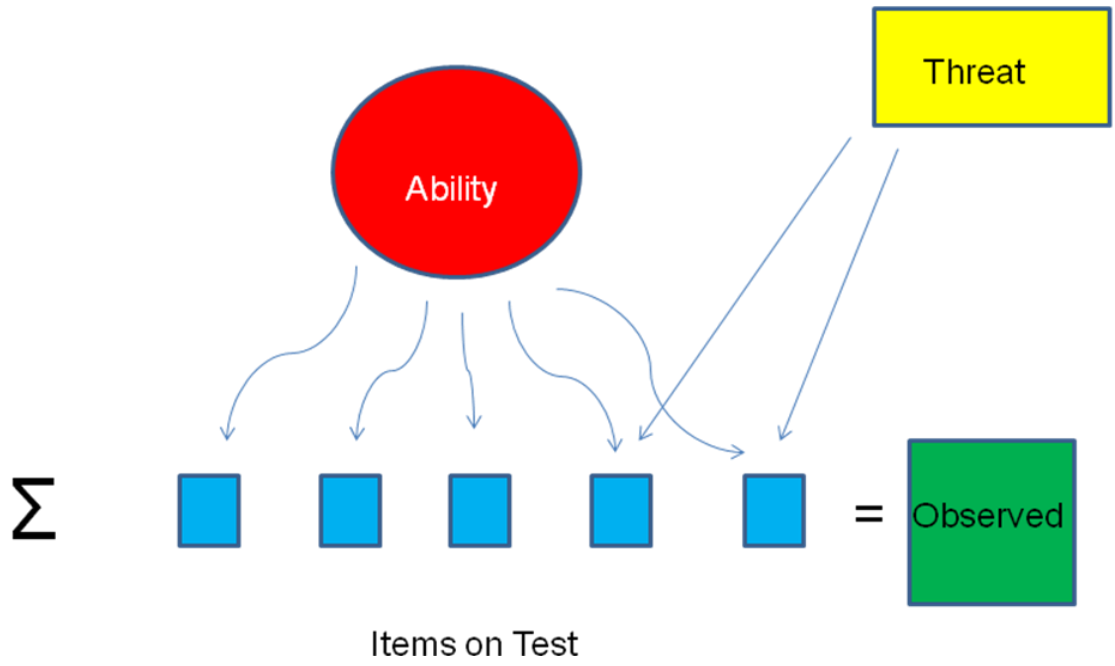

Figure 1. Item level impact of stereotype threat. A visual representation of the use of item analysis in assessing the contribution of individual items to overall test performance.

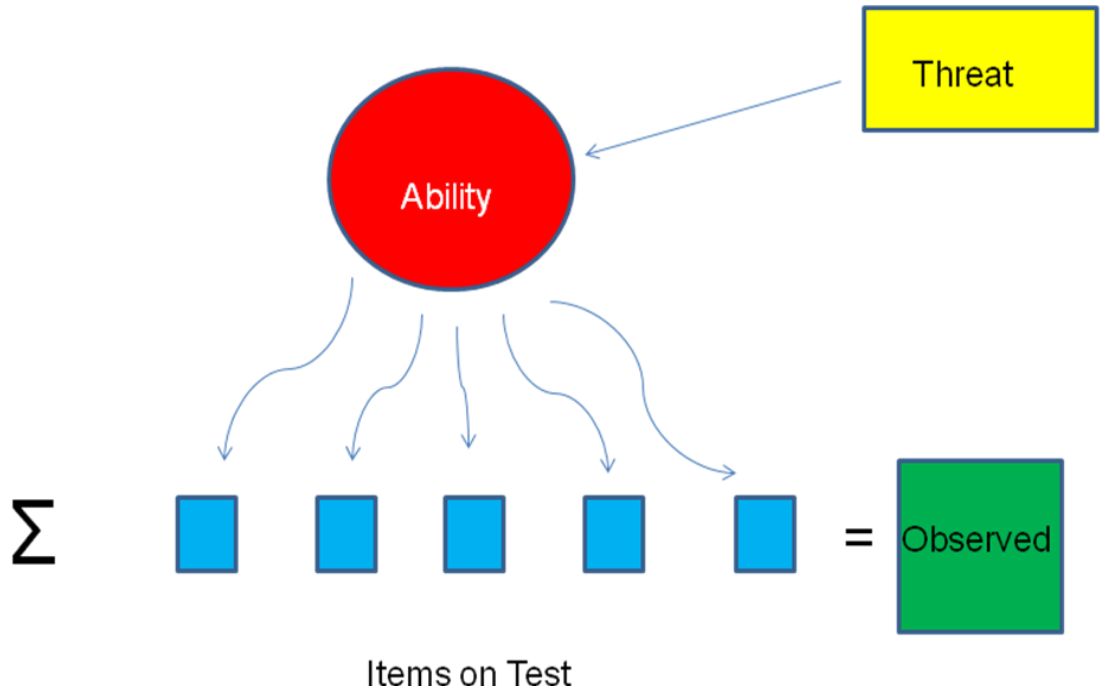

Figure 2. Ability level impact of stereotype threat. A visual representation of the use of item analysis in assessing the contribution of individual ability on overall test performance.

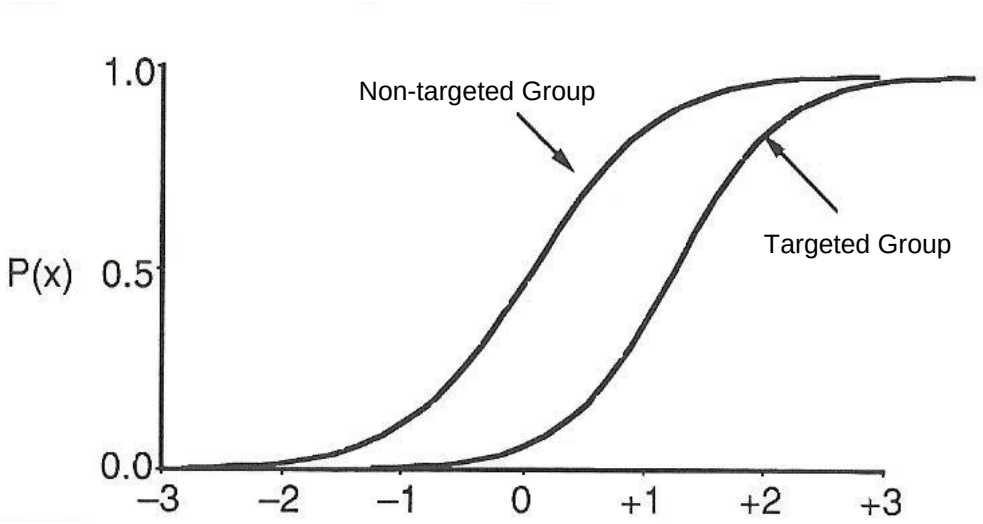

Figure 3. Uniform differential item functioning. An example of uniform differential item functioning, in which the difference between the targeted and the non-targeted groups is consistent across all possible values of theta.

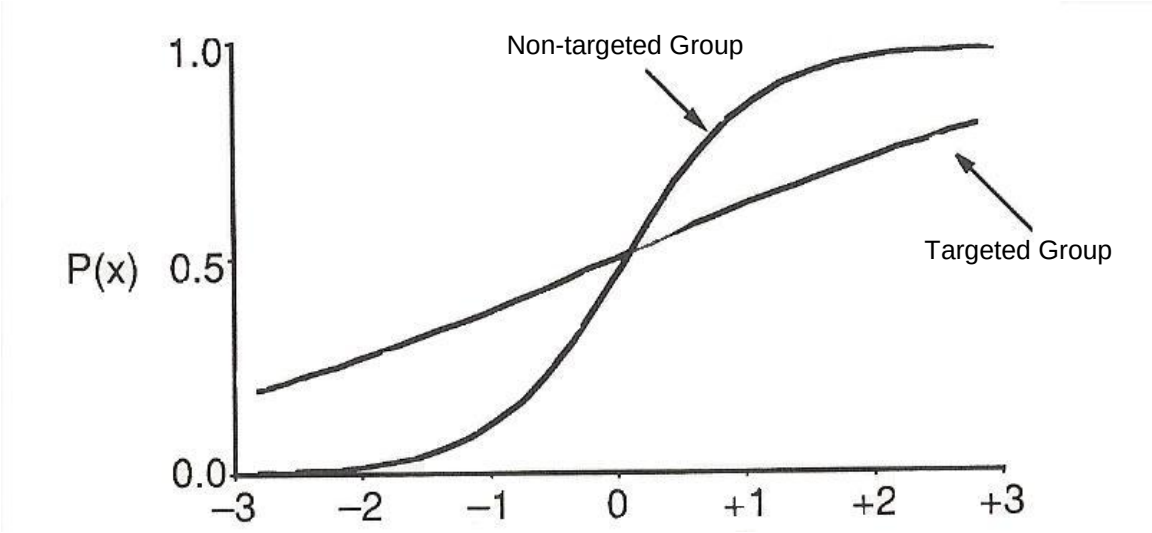

Figure 4. Non-uniform differential item functioning. An example of non-uniform differential item functioning, in which the difference between the targeted and the non-targeted groups is not consistent across all possible values of theta.

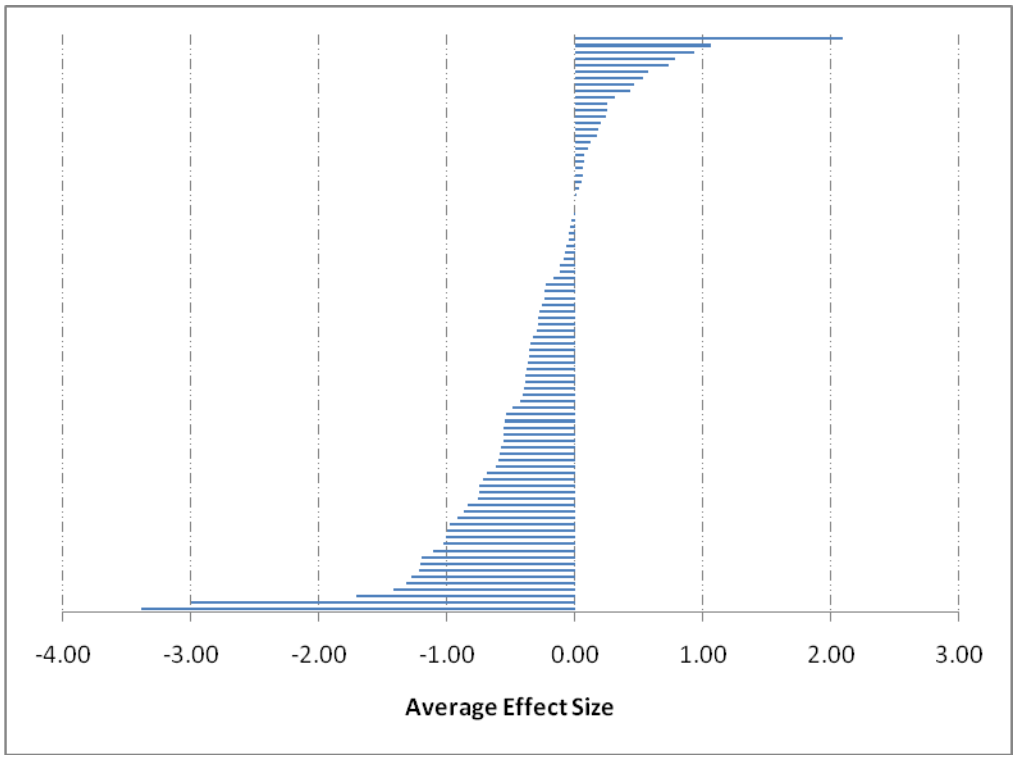

Figure 5. Distribution of effect sizes. The two distinctive distributions that arise from reviewing the meta-analytic effect sizes: one that represents stereotype threat and one that represents stereotype lift.

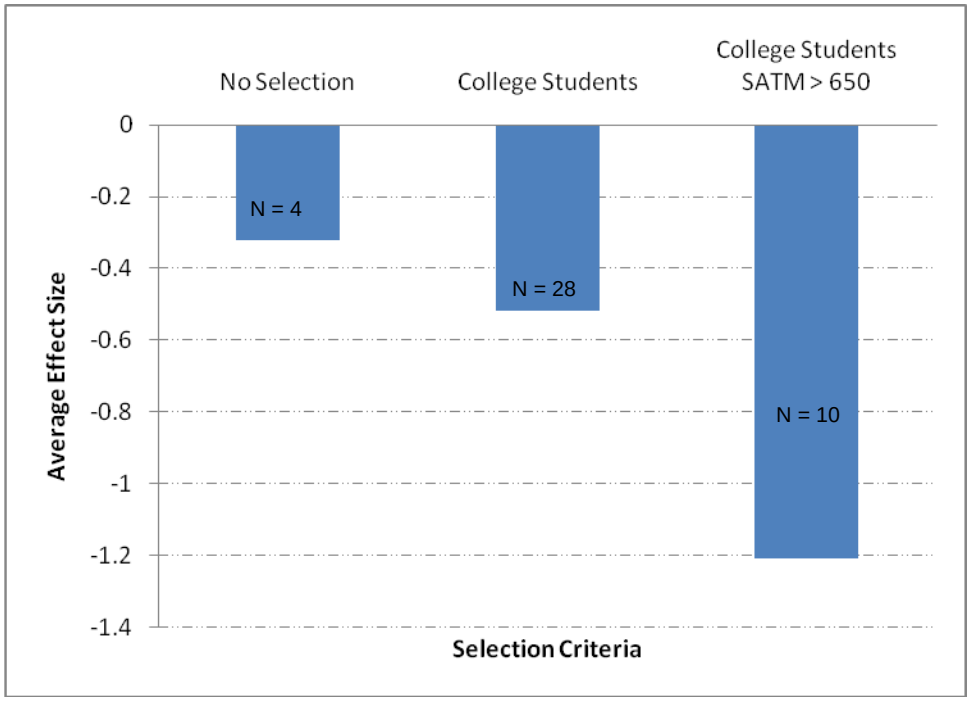

Figure 6. Average effect size by participant selection. The impact of participant selection on the effect size estimate of stereotype threat.

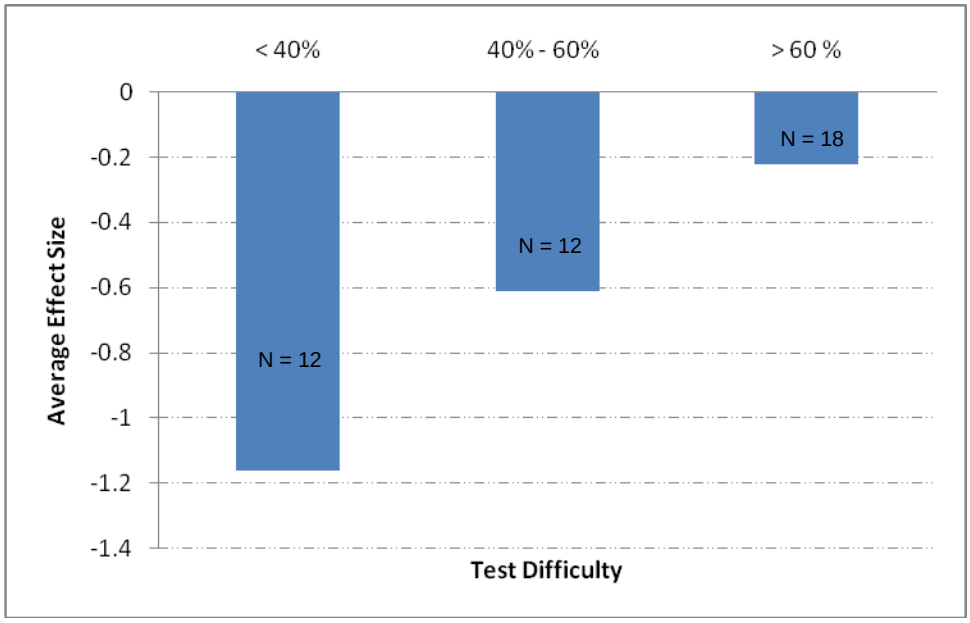

Figure 7. Average effect size by test difficulty. The impact of test difficulty on the effect size estimate of stereotype threat.

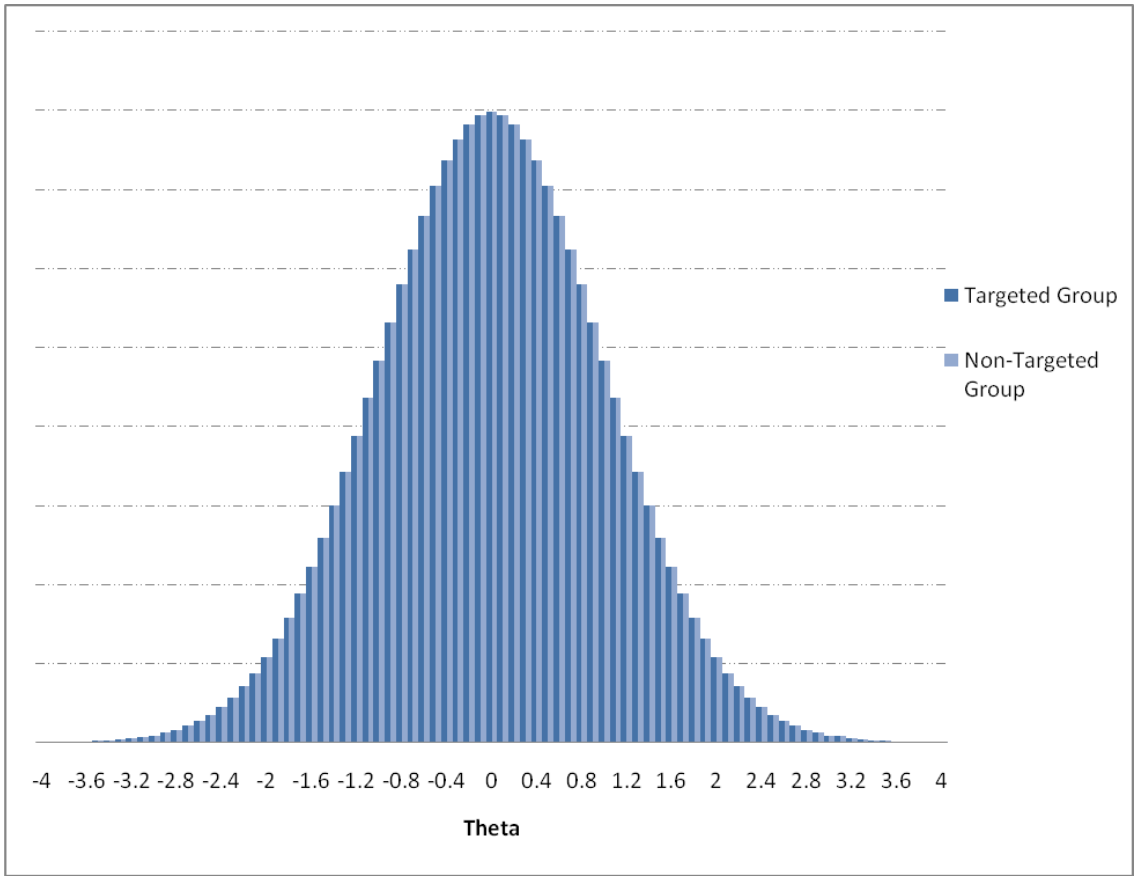

Figure 8. Equal target and non-target group theoretical distributions. Distribution of targeted and non-targeted participants created for the sensitivity analysis under the assumption that there is no true difference in latent proficiency in the target domain between the two groups.

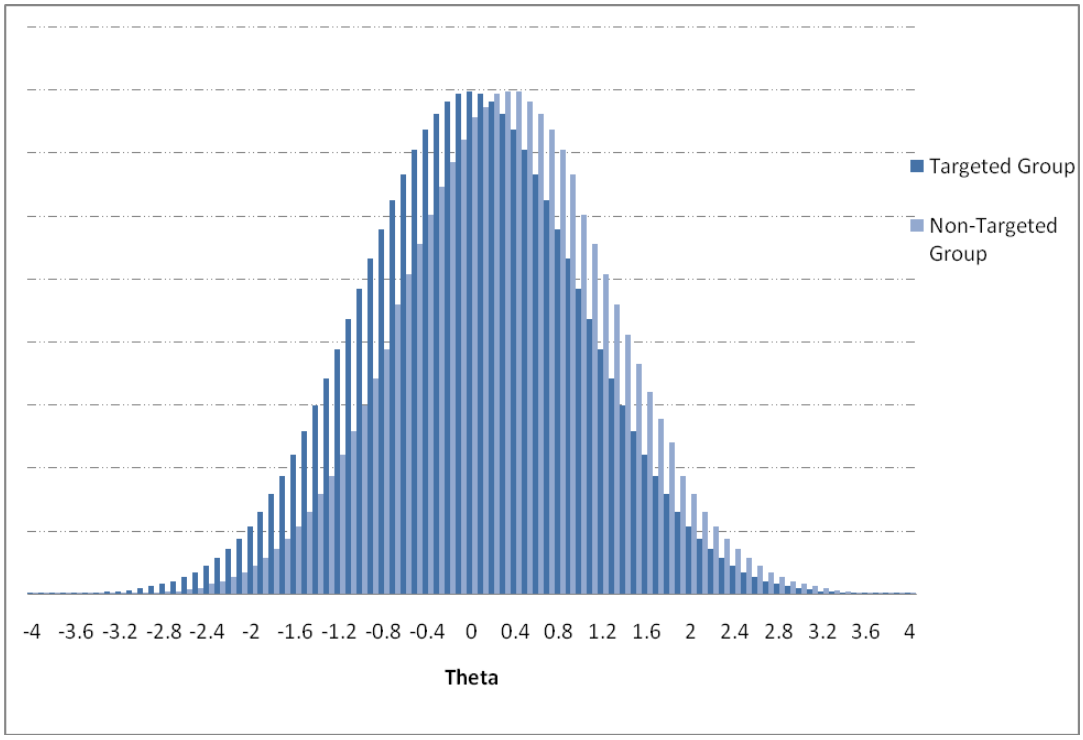

Figure 9. Unequal target and non-target group theoretical distributions. Distribution of targeted and non-targeted participants created for the sensitivity analysis under the assumption that there is a difference in latent proficiency in the target domain between the two groups.

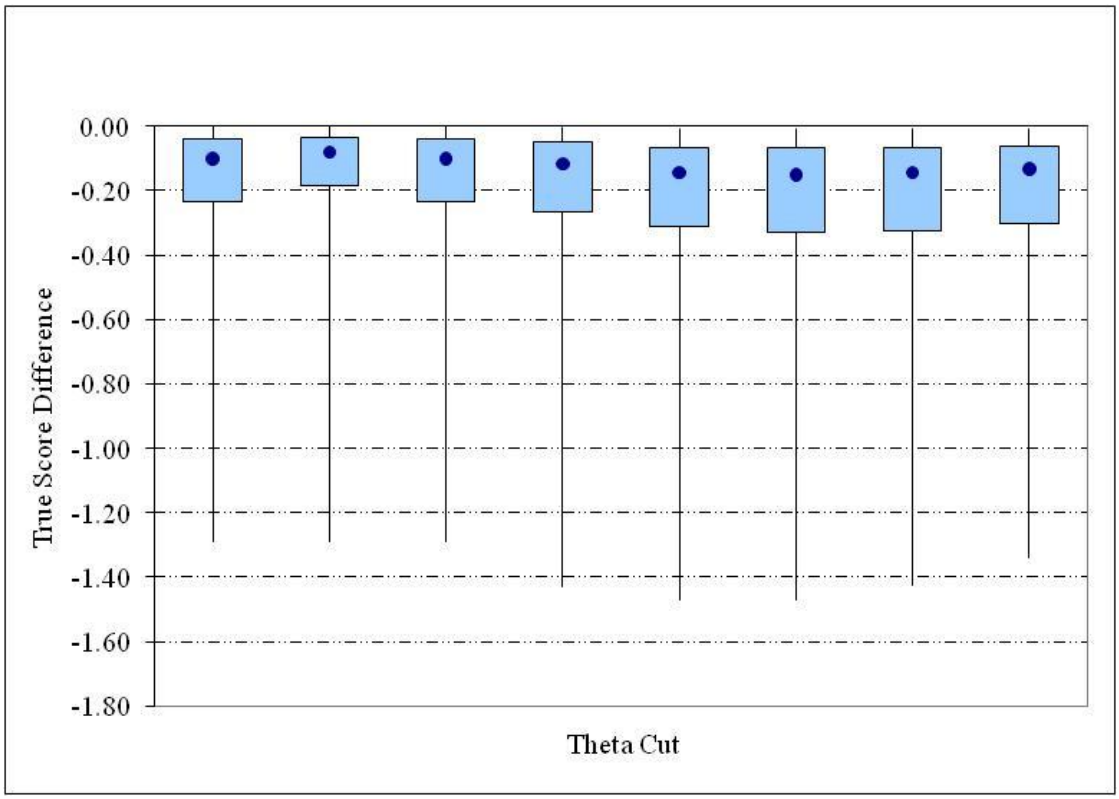

Figure 10. True score difference assuming equal groups. Difference in raw score across theta cut assuming equal groups while varying all four GAPS parameters.

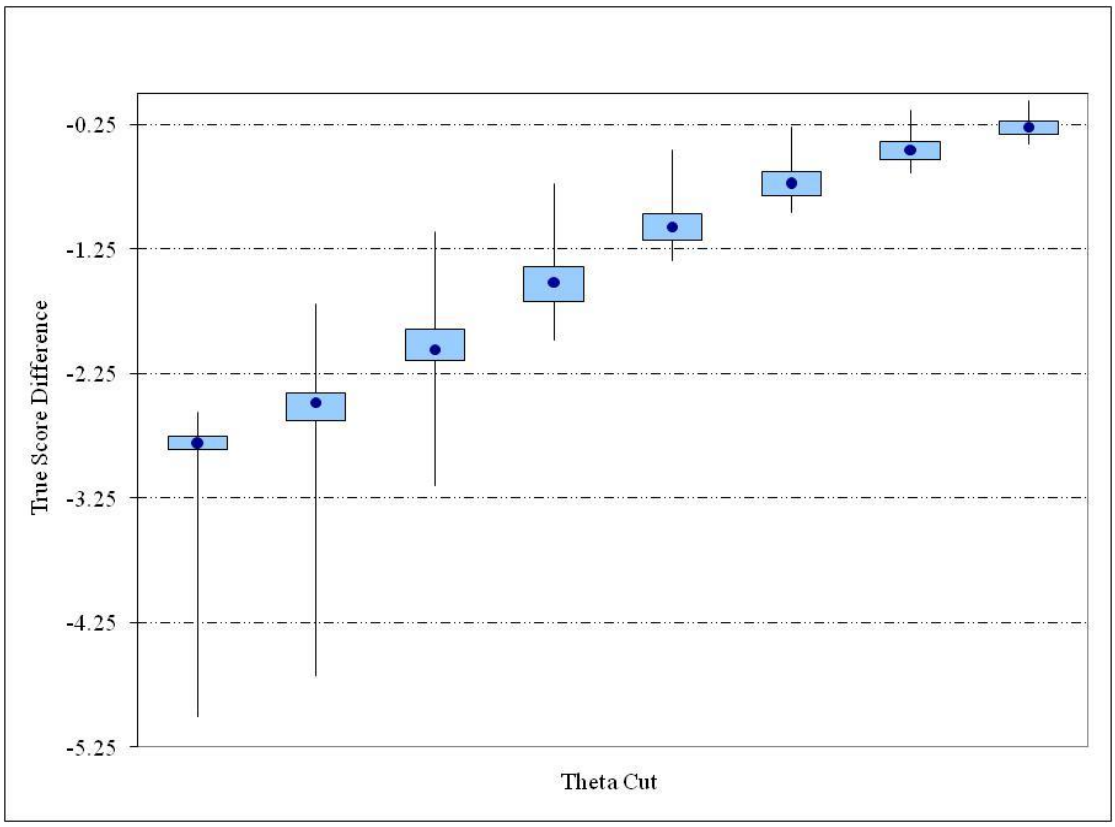

Figure 11. True score difference assuming unequal groups. Difference in raw score across theta cut assuming unequal groups while varying all four GAPS parameters.

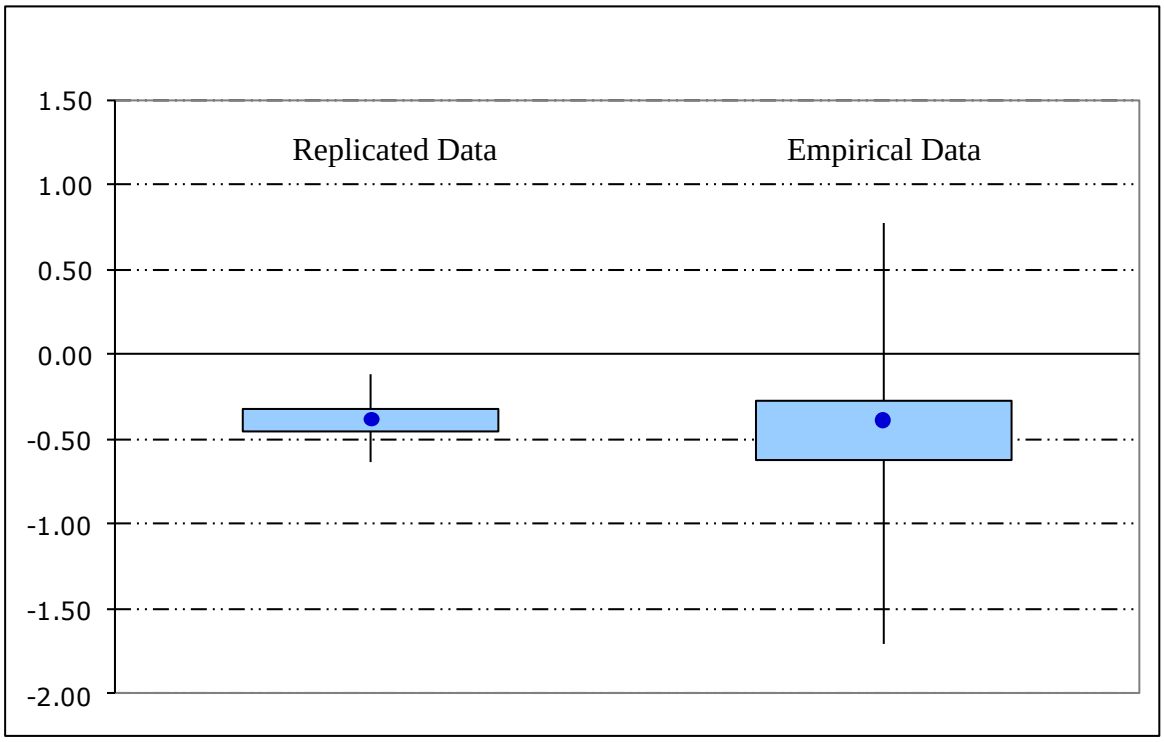

Figure 12. Boxplot of replicated and empirical effect sizes. Comparison of distributions of simulated stereotype threat studies and empirical distribution obtained from meta-analytic data.