Abstract

Previous studies of differential item functioning (DIF) have involved mostly analyses of cognitive ability measures and have produced little evidence that such items are biased in the sense that equally able individuals from different groups exhibit differing probabilities of selecting correct item responses. We tested hypotheses about DIF on a situational judgment test based on the notion that African American and Caucasian students have different opportunities to experience successful and unsuccessful solutions to common academic situations. Some support for the hypothesis that minority students might be better able to identify the worst solutions to situational judgment items was found, but there were no differences in the number of "biased" items when respondents were asked to identify optimal solutions to situational judgment items.

Interpreting Differential Item Functioning in a Situational Judgment Test: A Matter of Differential Access to Opportunities?

Racial group differences that arise in commonly used personnel selection measures evoke concerns about accuracy and fairness in testing. The most widely researched of these is the disparity between African Americans and Caucasians on tests of cognitive ability. The mean difference between these two groups tends to be approximately one standard deviation in magnitude (Neisser et al., 1996; Roth, Bevier, Bobko, Switzer, & Tyler, 2001; Schmidt 2002). For measures of other constructs like personality, biodata, and integrity, or for methods like structured interviews, differences tend to be smaller but are often still practically significant (e.g., Bobko, Roth, & Potosky, 1999; Clevenger, Pereira, Wiechmann, Schmitt, & Harvey, 2001). Although their possible causes have been studied in the industrial/organizational literature and beyond, racial group differences on various tests remain difficult to explain (Nisbett, 2005; Rushton & Jensen, 2005).

When individuals are assumed to be equivalent with respect to the constructs measured by a test, differences in test performance by racial group membership may still arise due to measurement bias (Society for Industrial and Organizational Psychology, 2003). Because "measurement bias is more complicated and cannot be addressed adequately using simple statistical or classical test theory methods" (Stark, Chernyshenko, & Drasgow, 2004, p. 498), few studies have examined the possible causes of group differences in test performance. One statistical method for investigating measurement bias that has become more useful with advances in item response theory (IRT) and computing technology is the examination of differential item functioning (DIF). DIF occurs when an item operates differently for people a certain group based on some factor

other than that which the test purportedly measures. Thus, members of one group can achieve higher scores on an item, on average, than members of another group, even when people in both groups are matched to be at the same level of the attribute tested. Despite the advantages of using IRT, studies of DIF have rarely examined patterns of results in light of substantive theories about the possible causes of item-level bias (Whitney & Schmitt, 1997). Instead, studies have tended to use a purely statistical approach by eliminating biased items during test development or generating post hoc explanations of DIF using item characteristics (e.g., Cole, 1981; Scheuneman & Gerritz, 1990; Tatsuoka, Linn, Tatsuoka, & Yamamoto, 1988). Because such approaches rely heavily on significance tests that are influenced by sample size, there exists a need for guidelines about how to make practically meaningful interpretations of DIF results, as Stark et al. (2004) point out.

The current study extends the application of traditional DIF detection procedures in two primary ways in the hopes of facilitating the testing of more meaningful, theory-driven hypotheses about group differences. First, we provide one of the first examinations of DIF within a test of situational judgment. Unlike investigations of DIF in cognitively based measures or tests of relatively stable personality traits, situational judgment test scores appear to be determined, at least in part, by knowledge and skills gained through experience (McDaniel & Nguyen, 2001). Second, we move beyond exploratory DIF analyses by hypothesizing *a priori* a specific pattern of DIF in the situational judgment test based on the theoretical assumption that disadvantaged racial groups have reduced access to developmental and achievement experiences. The

interpretation of results based on this approach is provided in contrast to interpretations based on a more traditional analysis of DIF.

Racial Group Differences Affecting Test Performance

Although many explanations for differences in cognitive ability test scores have been proposed and researched extensively (see Rushton & Jensen, 2005; Schmidt 2002), examinations of the causes of measurement bias have been scarce, particularly for noncognitively-based measures. Yet, one popular explanation for racial group differences is commonly espoused by researchers and the public alike. The simple, reasonable notion is that racial minorities are disadvantaged with respect to test performance (e.g., for personnel selection measures) because they have limited access to developmental and achievement experiences (Jensen, 1999; Neisser et al., 1996).

Whether the result of racial discrimination (e.g., Ogbu, 1978, 1994), cultural disadvantages (e.g., Freedle & Kostin, 1990; Scarr, 1994), or the relationship between minority status and lower socioeconomic status (e.g., Rushton & Jensen, 2005; Scheuneman & Gerritz, 1990), *the differential opportunities hypothesis* predicts that minority groups who have been denied relevant developmental opportunities will tend to perform worse than White individuals (Deutsch & Brown, 1964; Jachuck & Mohanty, 1974). At the same time, the hypothesis assumes that groups are endowed with the same level of different latent abilities; all people are equal at the outset. Thus, the main reason for group differences stems from the fact that the environment consistently hinders disadvantaged minorities from demonstrating fully their underlying ability, or, conversely, that the environment enhances majority group members' expressions of their ability. Stated another way, all people will develop at roughly equal rates (i.e., everyone

will acquire knowledge about the world), but only some groups will have access to certain types of experiences and development.

One notable line of research that is closely related to this notion of differential opportunities is that on Jensen's (1966, 1974) "cumulative deficit" hypothesis. Over time, constant environmental deprivation leads disadvantaged groups to fall increasingly behind in development and performance, a phenomenon labeled the cumulative deficit. The theory has provided the impetus for well known interventions aimed at compensatory education such as Head Start (Jensen, 1974). Unfortunately, studies (e.g., Campbell, Ramey, Pungello, Sparling, & Miller-Johnson, 2002; Cox, 1983; Jensen, 1977; Lazar & Darlington, 1982; Rayder, Body, & Nimnicht, 1978) have often failed to support strongly or consistently the cumulative deficit hypothesis for African American-Caucasian differences in cognitive ability (Jensen, 1999), leaving researchers uncertain about the environmental influences that might affect group differences.

Despite the lack of evidence supporting a cumulative deficit explanation, there are still good reasons for believing that the acquisition of specific skills and knowledge may be improved or hindered by environmental factors. First, Neisser et al. (1996) concluded that environment may still affect ability development despite the general lack of evidence for any specific environmental factor (Hernstein & Murray, 1995). They pointed out that, in recent decades, there have been substantial gains in academic achievement for underprivileged racial minorities (also see Gottfredson, 2005; Nisbett, 2005) but not for Whites, and that the Flynn effect, a steady increase in population IQ scores over time, may result from other environmental influences in the aggregate (Neisser, 1998). Second, the differential opportunities hypothesis (unlike the cumulative deficit

hypothesis) explains why one would see group differences with characteristics and achievements that are considered to be more malleable over time than cognitive ability (see Rushton & Jensen, 2005). Clearly, gains in knowledge and skills can and do occur, as is often demonstrated in the training and educational research literature (e.g., Carretta & Ree, 2000). Hence, we propose that people with equivalent levels of latent ability may perform differently on tests of experience-based knowledge and skills when some are afforded access to certain developmental opportunities while others have limited access to those same opportunities. Although this proposition is rather broad, we describe a unique application of this theory to situational judgment test performance that is both specific and substantive in nature.

Situational Judgment Tests and Experience

Situational judgment tests (SJTs), sometimes viewed as low-fidelity simulations (Motowidlo, Dunnette, & Carter, 1990; Motowidlo, Hanson, & Crafts, 1997), consist of a set of hypothetical dilemmas and corresponding options describing how one might typically react in response to a given dilemma (McDaniel & Nguyen, 2001). Although they have been developed in alternative formats (e.g., video-based; Chan & Schmitt, 1997), SJTs are typically administered in paper-and-pencil multiple-choice formats. SJTs have become popular in selection (Peeters & Lievens, 2005) probably because they have demonstrated practically significant criterion-related validities (ρ = .34; McDaniel, Bruhn-Finnegan, Morgeson, Campion, & Braverman, 2001), produce lower adverse impact on minority groups (Chan & Schmitt, 1997; Nguyen & McDaniel, 2003; Pulakos & Schmitt, 1996; Weekley & Jones, 1999), and have good face validity.

Over the years, researchers have questioned what exactly it is that SJTs measure (McDaniel et al., 2001). Some have proposed that "situational judgment" represents a specific construct, perhaps akin to intelligence, while others have viewed SJTs as a generic method that can be tailored to measure different types of core constructs such as personality, values, cognitive ability, and knowledge (see McDaniel et al., 2001; Schmitt & Chan, in press). Because the dilemmas posed in SJTs may take place in different domains (e.g., on-task at a job, in school, at home with friends) and because items may lack an objectively correct answer, it seems that SJTs can measure multiple constructs if constructed with that aim. Yet, Schmitt and Chan (in press) recently reviewed the SJT literature related to measurement of known psychological constructs and concluded that SJTs seem to measure not only traditional constructs but also some unique attribute related to adaptability and practical intelligence.

If one accepts that SJTs can at least measure some characteristics beyond general cognitive ability, it is certainly possible that experience in and knowledge of real-life situations can help test takers determine the merit of possible courses of action (McDaniel & Nguyen, 2001). Even SJTs that are weakly correlated with cognitive ability could measure a basic form of trial-and-error learning (i.e., someone was in a particular situation and learned about whether or not a solution strategy worked). SJT scores have been correlated with age and experience (e.g., Smith & McDaniel, 1998; Weekley & Jones, 1999). In addition, they look and act very much like Wagner and Sternberg's (1991) Tacit Knowledge Inventory for Managers (McDaniel et al., 2001; Motowidlo et al., 1990), and "tacit knowledge" is theorized to develop through experience

(Sternberg et al., 2000). Therefore, the skill(s) or knowledge measured by an SJT can be examined in the context of racial group differences caused by differences in experience.

One feature of many SJTs, the response format, makes them particularly suitable for a simple test of the differential opportunities hypothesis. Although response formats may vary in wording (e.g., asking what one would do vs. what is the best action to take), each item typically asks respondents to indicate the action that they are *most* likely to perform and the action that they are *least* likely to perform in a given situation. For SJTs scored with the method developed by Motwidlo et al. (1990) and Motowidlo, Russell, Carter, & Dunnette (1988), each item score is a composite index based on respondents' answers to these two questions for each item, or dilemma. Specifically, the responses to these two questions are compared to experts' ratings in such a way that a respondent receives +1 point for matching the most likely response to the best rated option, no points for a neutral option, and -1 point for selecting the worst rated option. Answers to the least likely question are scored in the opposite manner. The most and least component scores are then summed to form an item-level score (ranging from +2 to -2). Conceptually, the item-level scores indicate a person's level of situational judgment, though the value (e.g., +1) does not correspond directly to any particular response option on the SJT. Finally, the item scores are aggregated to form a test score.

If one views SJTs in light of the differential opportunities hypothesis, it seems that individuals with equal levels of situational judgment should perform differently on the test when the types of experiential knowledge and skills they have acquired are related to group membership. If racial minorities are at a disadvantage, they should acquire situational skills or knowledge just like any other persons, but only knowledge

relevant to certain types of situations (i.e., situations in which a person has limited opportunities for success). In these situations where development and achievement is limited, one would expect minorities to learn which situational responses are worse than others, even though none of their responses might be "effective" solutions on the spectrum of possible solution strategies.

We illustrate the point with a hypothetical example related to racial discrimination. Imagine a high school in which racial minority students are rarely assigned to be the group leader for class projects, simply because the teachers discriminate by race. When conflicts arise in project groups, a minority student can attempt to use various strategies for resolving the situation (e.g., asking peers politely to compromise, telling the teacher about the problem, or keeping silent and doing extra project work). However, the strategies available to the minority may all be relatively ineffective because the she has no leadership power to execute alternative solutions. As a result, the minority learns about the differences between ineffective situational responses and learns which strategies produce the worst outcomes, but does not acquire experience implementing the more effective strategies. Based on this logic, we predict that SJT scores referring to the *least* question (i.e., which strategy a person will most likely avoid) will show measurement bias in favor of racial minorities because they require a person to distinguish the worst response from other poor responses.

Turning to the majority group in this example, effective conflict resolution strategies that relate to leadership are only available to the majority group students. Being advantaged, majority students are placed in the role of leader more often and gain more experience. As group leaders, they are more likely to acquire a better

understanding of how different leadership strategies work (e.g., taking charge and delegating, or arbitrating the conflict) and produce positive group outcomes. Therefore, the majority group members in this high school gain more experiences related to leadership that help them distinguish between the effective solution strategies even though they do not gain more overall experience in project teams than the minorities. This logic complements that provided for the minority group; SJT scores referring to the *most* question (i.e., what should people do to produce the best outcome) will place majority group members at an advantage (and minorities at a disadvantage) because they have gained more experience relevant to effective situational responses.

While the real world clearly does not operate according to such simple principles as those presented in the example, commonly espoused claims about racial group differences resulting from racial discrimination at the societal level would be consistent with these general lines of reasoning. As a result, three specific hypotheses about item bias for SJT scores were formed:

> Hypothesis 1: SJT item scores for the *least* response will be biased in favor of racial minority group members who are of equal ability as majority group members.
>
> Hypothesis 2: SJT item scores for the *most* response will be biased in favor of racial majority group members who are of equal ability as minority group members.
>
> Hypothesis 3: SJT item scores based on a composite will show less bias than item scores based on a component score (i.e., answers to *least* or *most* responses) because of the two contradictory effects in Hypotheses 1 and 2.

These hypotheses were tested using patterns of differential item functioning with item response theory.

Differential Item Functioning

Based on item response theory (Lord, 1980; Lord & Novick, 1968), differential item functioning is the phenomenon that occurs when test items fail to assess equivalent individuals of different groups in the same manner (Dorans & Holland, 1993; Raju & Ellis, 2002). More specifically, an item shows DIF when individuals from different groups show different probabilities of selecting each answer choice, after they have been matched to have the same standing on the latent attribute assessed by the item. Because item scores would be expected to differ when one group simply has more people of higher ability (i.e., an explanation related to the distribution of the attribute; Rushton & Jensen, 2005), the matching of individuals on a test attribute means that DIF indices indicate item-level bias due to some cause other than the attribute (Dorans & Holland, 1993; Stark et al., 2004).

Typically, practitioners wish to identify bias in an item or test and eliminate it. Researchers sometimes investigate the matter further by searching rationally for similarity between the DIF items or persons (e.g., Robie, Zickar, & Schmit, 2001) to identify a substantive cause of differential response patterns (Whitney & Schmitt, 1997). However, such post hoc exploratory efforts tend to produce inconsistent findings (e.g., Cole, 1981; Scheuneman & Gerritz, 1990; Tatsuoka et al., 1988). Also, the early research literature on DIF usually focused on tests of cognitive ability (Saad & Sackett, 2002). As cognitive ability is a relatively stable attribute and ability tests have objectively correct answers, prior research has provided limited insight into the measurement and potential

causes of group differences in typical selection measures that are less ability laden. This study provides meaningful contributions along both of these fronts. In summary, we move beyond traditional examinations of race differences by predicting specific patterns of DIF in an SJT based on substantive theories about the differences in types of experiences acquired by people of different racial groups, who may possess the same levels of a latent attribute underlying test performance.

Method

Sample

Data were collected from 503 college freshmen during their first semester. Students were recruited from two large, Midwestern universities and paid $20 for participating. Although the study was open to all freshmen, African-Americans and Caucasians were targeted during recruitment through a variety of means including student newspapers, flyers in dormitories, registration tables outside of cafeterias, extracurricular groups, and (at one university) an email list of all African-American freshmen provided by the university registrar. Measures were administered in two-hour sessions, with no more than twenty-five participants in one session. (We administered a number of other measures during these sessions that are not relevant to the present study.) After excluding participants of other racial groups and screening for careless responders, usable data were obtained from 405 participants (230 Caucasians and 175 African-Americans). All participants were either 18 (N=309) or 19 years old (N=97), and 39.8% were male.

Measures

SJT. The 73 items administered in this study are a subset of items from a larger inventory, the Life Events Assessment and Development (LEAD), created for a prior data collection. The process used to develop the LEAD, including the generation of items and scoring key, is described in detail in Oswald, Schmitt, Kim, Ramsay, and Gillespie (2004). In brief, the SJT was specifically designed to measure students' judgment in a broad array of college-relevant situations within twelve academic and nonacademic areas valued by U. S. colleges (see Table 1 for a description of the areas) For each item, respondents were asked: "What are you *most* likely to do?" and "What are you *least* likely to do?" We developed a scoring key based on the responses of subject matter experts (i.e., junior- and senior-level students) according to rules in Motowidlo et al. (1990). Each item was then assigned a score ranging from -2 to +2 based on a participant's response to both the most and least likely questions. Sample items are presented in Appendix A.

Standard models used to analyze DIF rely on a strong assumption of unidimensionality (Embretson & Reise, 2000; Reckase, 1979). An exploratory factor analysis of the 73 items supported a two-factor solution. The factors, however, were not interpretable based on item content or structure. In addition, results clearly indicated that many items loaded equally well on both the first and second factors. Consequently, the scale was refined by removing items that loaded disproportionately on one factor and that had lower correlations with the rest of the items based on the factor loadings.

Subsequent analyses of the remaining 41 items produced results that supported a one-factor solution based on criteria for establishing unidimensionality recommended in previous work (Drasgow & Lissak, 1983; Embretson & Riese, 2000; Hattie, 1985; Reckase, 1979; University of Illinois IRT Modeling, 2005). The first eigenvalue was

dominant, accounting for 13.0% of the variance and being approximately 3.4 times greater than the second eigenvalue, and the correlation between the first and second factor was .79 when the items were forced onto two factors with oblique rotation. Also, the scree plot indicated a one-factor solution. Consequently, this modified scale was deemed to be sufficiently unidimensional so as to be suitable for IRT analyses. The alpha reliability estimate for the resulting scale was .81.

Demographics. Demographic information was also collected, including age, gender, ethnicity, citizenship status, and year in school.

Analyses

Three investigations of uniform DIF were conducted based on the dichotomous component score for the *most* response, the dichotomous score for the *least* response, and the polytomous composite score. All analyses were run in Parscale 4.1 (Muraki & Bock, 2003) using the 1-parameter latent trait, partial-credit model. This model provides a *location* parameter for a logistic curve to indicate the probability that a person will achieve a particular item score based on his or her latent trait level. Two items were excluded from these analyses because they did not satisfy Parscale's computational requirement every response option is selected by at least one person from each group.

The detection of DIF is more meaningful when the difference between location parameters between groups is statistically significant (Raju & Ellis, 2002) and when an item's χ^2 fit statistic is acceptable (Embretson & Reise, 2000). As a result, we used Lord's chi-square statistic to identify statistically significant DIF ($p<.05$). Items with a smaller location parameter are said to favor one group (i.e., require less ability to achieve a high score) over the other. Also, we removed items from *most* and *least* analyses when they

failed to achieve acceptable levels of fit for both racial subgroups (see Table 2). None of these items displayed DIF prior to their removal. Because Hypothesis 3 involves a comparison between analyses and because no items failed to fit both groups for most and least scores, all items were retained in the composite score analysis. In the end, the exclusion of items in each analysis did not significantly alter the unidimensionality of the scale.

Results

Regarding descriptive statistics, mean scores on the 39 SJT items analyzed were .55 (*SD*=.35) for Caucasians and .67 (*SD*=.36) for African Americans, with the standardized mean difference, *d*, equal to .34. (The African American group also had a higher mean score on the 32 items excluded from the IRT analyses, with *d*=.11.) Although the LEAD was not expected to produce a higher group mean score for African Americans, DIF is related to the mean difference on the composite score only indirectly (Stark et al., 2004), and this finding does not obviate the need for more refined analyses of DIF. No statistically significant differences in the SJT were found for students' university membership. However, males in this sample scored considerably lower on the SJT than females (*d*=.63).

The DIF results across the three analyses are included in Table 2. The pattern of statistically significant DIF provides support for Hypothesis 1. Seven items showed DIF based on the *least* responses, with five favoring African Americans. The pattern predicted by Hypothesis 2, however, was not supported. Of the six items showing DIF based on responses to the *most* question, three favored each group. Only one item produced

evidence of DIF for both most and least (and the composite) responses, and Caucasians were favored in both instances.

As for the composite item scores, four items showed statistically significant DIF. Three of these favored Caucasians, again indicating that the mean difference in overall scores is not a direct function of DIF. Although the difference in number of items showing DIF between analyses of the *least*, *most*, and composite scores is not significantly different statistically, an examination of the differences in location parameters across all items in the three analyses indicates some support for Hypothesis 3. While the most responses did not produce DIF that favored Caucasians on the whole, it seems that bias towards one group for a *most* response is counterbalanced by the *least* response in most cases.

Following more traditional approaches, we also examined the patterns of DIF post hoc, as they relate to the 12 content areas described in Table 1. Unfortunately, the number of items per content area ranged from 0 to 6 as a result of the scale-refinement procedures (mentioned earlier) to preserve unidimensionality. Table 3 shows the DIF items found within each content area. None of the six Knowledge items or three Social Responsibility items displayed DIF, nor did the single Continuous Learning or single Adaptability item. Also, just 1 of 6 Perseverance items showed DIF. In contrast, 3 of 6 Interpersonal Skills items and 3 of 5 Career Orientation items showed DIF, though neither racial group was clearly favored. It then seems that items referring to Career Orientation and Interpersonal Skills dilemmas are particularly susceptible to DIF. As these statements are obviously based on a small number of items per content area, further investigation of certain domain-specific situations may be worthwhile.

Discussion

We examined DIF in an SJT designed to assess college student performance in multiple domains using the differential opportunities hypothesis to formulate three a priori hypotheses about DIF. Based on the assumption that members of different racial groups acquire different types of experiences based on cultural or societal advantages/disadvantages, we predicted that African American students would generally achieve better scores than Caucasians when asked to distinguish between *ineffective* courses of action on the SJT, even after individuals are matched on the dominant attribute measured by the test. The results produced some support for this prediction as items scored according to least responses were more often (5 vs. 2) biased in favor of African Americans. A complementary prediction for Caucasians in the opposite direction, however, was not supported. We also found that examining the *most* and *least* component scores produced different patterns of DIF than did the composite score.

One reasonable explanation for the finding that the component SJT scores produced substantial DIF while the composite scores did not is that DIF can be masked at the test level (Roznowski, 1987) when items favoring one group generally cancel out items favoring another group to form an unbiased test (Raju & Ellis, 2002; Stark et al., 2004). Although there are important theoretical and practical distinctions between a test free of DIF items and a test in which DIF items balance each other, the result of the tests' predictive function may be the same in either scenario.

Finally, it is important to note that our SJT, the LEAD, is only weakly correlated with cognitive ability (as measured by the SAT and ACT college admissions tests). Items in Oswald et al. (2004) were correlated -.03 with cognitive ability, the 71 SJT items for

which data were collected in this study were correlated -.06, and the 39 items retained for analyses were correlated -.10. Although an SJT's cognitive loading may help to explain racial group differences in general (e.g., Nguyen & McDaniel, 2003), differences in cognitive ability cannot be used to explain the patterns of DIF found here in the LEAD SJT.

Limitations and Future Directions

Limitations of this study center around two issues: statistical significance and the manner in which an SJT can be constructed and scored. As in all studies of DIF, we initially relied on significance tests for detecting DIF. Deciding whether a particular item operates differently "enough" to be exhibiting DIF in two samples is somewhat subjective. We reduced the role of chance by examining DIF in light of a priori hypotheses but examined the results in a dichotomous fashion, whereby items either did or did not show DIF. Nonetheless, researchers must consider statistical power as it relates to sample size, the number of items studied, and the number of levels produced by the response options when interpreting DIF results.

Regarding the second limitation, confusion about what SJTs measure in general poses a potential problem for studying DIF. Our SJT was constructed for the purpose of measuring students' abilities to cope with future dilemmas in a college context across a wide range of settings. We cannot say with absolute certainty that the ability determining respondents' scores was controlled for with the IRT models used. However, we did find sufficient psychometric evidence of unidimensionality to justify using the one-parameter IRT model. DIF methods are relatively robust to violations of strict unidimensionality when the first factor is dominant (Reckase, 1979) and when first-order factors are highly

correlated with each other (Drasgow & Parsons, 1983). As both of these criteria were met by our SJT, any effect of multidimensionality on the results is probably attenuated. Even so, the theory-based approach to DIF used here could be applied to purer measures of constructs (perhaps personality or biodata) and linked to other measures of access to opportunities afforded to specific racial groups.

Another challenge to our interpretations of DIF lies in the fact that the most and least component responses are not independent of the composite item scores (McDaniel & Nguyen, 2001). The two models (i.e., dichotomous and polytomous) used to analyze DIF create a confound that might account for differences in DIF patterns across most, least, and component scores. The power to detect DIF in dichotomously scored items (i.e., most/least) is higher than in composite scores because fewer parameters need to be estimated. At the same time, composite items are scored on a continuum and thus provide more information, though there are psychometric issues of partial ipsativity (Hicks, 1970). At the present, the literature does not provide much guidance about how methodological differences such as these would affect practical interpretations of DIF (Zickar, 2002). Still, support for Hypothesis 3 appears to be explained better and more parsimoniously by the well known phenomenon of compensatory DIF (Raju & Ellis, 2002), whereby items favoring one group cancel out items favoring another group. While compensation clearly occurs at the test level after aggregating items, it seems that compensation can also occur at the item level when aggregating component scores based on answers to response options.

One interesting direction for future research suggested by our findings is the investigation of DIF in racial groups by gender. Unfortunately, the race x gender

subgroup sample sizes in this study are not large enough to warrant more focused analyses, though mean differences within these smaller subgroups were found (see Drzakowski et al., 2004). If future work replicates our gender-related findings, perhaps these subgroups should be examined separately based on more complex hypotheses. It is plausible that certain groups (e.g., female African Americans) are doubly disadvantaged, given the literature on such effects as the "glass ceiling" in managerial jobs (Federal Glass Ceiling Commission, 1995).

Another question evoked by the logic behind our research hypotheses is whether measurement bias on SJTs is affected by the racial group membership of the experts used to develop an empirical scoring key. Members of one group might believe that a particular response option is highly effective, whereas members of another group believe it is less effective, perhaps because it is not culturally valued or because it is not a realistic option for members of that particular group. For example, minority group members who lack role models may be less likely to consult with an academic advisor about academic or social problems than majority group members. In some initial work, Kim, Schmitt, Friede, Oswald, Ramsay, and Gillespie (2004) compared scoring keys based on Caucasian and African American students of equivalent academic status using the SJT described in Oswald et al. (2001). They found few differences with respect to the experts' ratings of effectiveness for the response options (during key development). They also found that keys based on majority and minority group experts produced similar patterns of uniform DIF but different patterns of nonuniform DIF (i.e., the discrimination parameter in an IRT model is different across groups; Raju & Ellis, 2002). Nonetheless, further in-depth studies may provide more insight on this issue.

Before concluding, we acknowledge that the differential opportunities hypothesis advanced in this paper is quite broad. Since we tested the hypothesis with just one type of measure and used race as a proxy for types of experience, we do not claim definitive conclusions about the causes of differences in test performance between Caucasians and African Americans. Rather, this study advances a new approach for studying DIF in a meaningful manner using theory-driven expectations. The results provide mixed support for the general hypothesis but suggest that patterns of DIF, when modeled appropriately, may reflect meaningful psychological phenomena.

References

Bobko, P., Roth, P. L., & Potosky, D. (1999). Derivation and implications of a meta-analytic matrix incorporating cognitive ability, alternative predictors, and job performance. *Personnel Psychology, 52*, 561-589.

Campbell, F. A., Ramey, C. T., Pungello, E., Sparling, J., & Miller-Johnson, S. (2002). Early childhood education: Young adult outcomes from the Abecedarian Project. *Applied Developmental Science, 6*, 42-57.

Carretta, T. R. & Ree, M. J. (2000). General and specific cognitive and psychomotor abilities in personnel selection: The prediction of training and job performance. *International Journal of Selection and Assessment, 8*, 227-236.

Chan, D. & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests. *Journal of Applied Psychology, 82*, 143-59.

Clevenger, J., Pereira, G. M., Wiechmann, D., Schmitt, N., & Harvey, V. S. (2001). Incremental validity of situational judgment tests. *Journal of Applied Psychology, 86*, 410-417.

Cole, N. S. (1981). Bias in testing. *American Psychologist, 36*, 1076-1077.

Cox, T. (1983). Cumulative deficit in culturally disadvantaged children. *British Journal of Educational Psychology, 53*, 317-326.

Deutsch, M. & Brown, B. (1964). Social influences in Negro-White intelligence differences. *Journal of Social Issues, 20*, 24-35.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardizatoin. In Holland, P. W. and Wainer H. (Eds.),

Differential Item Functioning. Hillsdale, NJ: Erlbaum.

Drasgow, F., & Lissak, R. I. (1983). Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. *Journal of Applied Psychology, 68*, 363-373.

Drasgow, F. & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Pscyhological Measurement, 7*, 189-199.

Drzakowski, S., Friede, A., Imus, A., Kim, B., Oswald, F., Schmitt, N., & Shivpuri, S. (2004). *Report: Further development of the Assessment of Life Experiences Questionnaire (ALEQ) biodata measure and Life Events Assessment and Development (LEAD) situational judgment inventory: Investigating differences across gender, ethnic and institutional subgroups*. Princeton, NJ: College Board.

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Federal Glass Ceiling Commission (1995). *A solid investment: Making full use of the nation's human capital. Recommendations of the Federal Glass Ceiling Commission*. Washington D. C.: Author.

Freedle, R., & Kostin, I. (1990). Item difficulty of four verbal item types and an index of Differential Item Functioning for Black and White examinees. *Journal of Educational Measurement, 27*, 329-343.

Gottfredson, L. S. (2005) "What if the hereditarian hypothesis is true". *Psychology, Public Policy, and Law 11*, 311–319.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*, 139-164.

Herrnstein, R. J., & Murray, R. (1995). *The bell curve: Intelligence and class structure in American life*. New York: Free Press.

Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin, 74*, 167-184.

Jachuck, K. & Mohanty, A. K. (1974). Low socio-economic status and progressive retardation in cognitive skills: A test of cumulative deficit hypothesis. *Indian Journal of Mental Retardation, 7*, 36-45.

Jensen, A. R. (1966). Cumulative deficit in compensatory education. *Journal of School Psychology, 4*, 37-47.

Jensen, A. R. (1974). Cumulative deficit: A testable hypothesis? *Developmental Psychology, 10*, 996-1019.

Jensen, A. R. (1977). Cumulative deficit in IQ of blacks in the rural south. *Developmental Psychology, 13*, 184-191.

Jensen, A. R. (1999). *The g factor: The science of mental ability.* Westport, CT: Praeger.

Kim, B., H., Schmitt, N., Friede, A., Oswald, F. L., Ramsay, L. J., & Gillespie, M. A. (2004). *Differential item functioning in situational judgment tests: Is it a function of the scoring procedure?* Paper presented at the 19th Annual Convention of the Society for Industrial and Organizational Psychology, Chicago, IL.

Lazar, I. & Darlington, R., Murray, H., Royce, J., & Snipper, A. (1982). Lasting effects of early education: A report from the Consortium for Longitudinal Studies. *Monographs of the Society for Research in Child Development, 47*, 1-141.

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

Lord, F.M. & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley.

McDaniel, M. A., Bruhn-Finnegan, E. B., Morgeson, F. P., Campion, M. A., & Braverman, E. P. (2001). Predicting job performance using situational judgment tests. *Journal of Applied Psychology, 86,*730-740.

McDaniel, M. A., & Nguyen, N. T. (2001). Situational judgment tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment, 9*, 103-113.

Motowidlo, S.J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology, 75*, 640-647.

Motowidlo, S. J., Hanson, M. A., & Crafts, J. L. (1997). Low-fidelity simulations. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement methods in industrial psychology*. Palo Alto, CA: Davies-Black Publishing.

Motowidlo, S. J., Russell, T. L., Carter, G. W., & Dunnette, M. D. (1988). *Revision of the Mangement Selection Interview: Final report*. Minneapolis, MN: Personnel Decisions Research Institute.

Muraki, E. & Bock, R. D. (2003). PARSCALE for windows (Version 4.1). Chicago: Scientific Software.

Neisser, U. (1998). *The rising curve: Long-term gains in IQ and related measures*. Washington, DC: American Psychological Association.

Neisser, U., Boodoo, G., Bouchard, T. J., Jr., Boykin, A. W., Brody, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R. J., & Urbina, S. (1996). *Intelligence: Knowns and unknowns. American Psychologist, 51*, 77-101.

Nisbett, R. E. (2005). Heredity, environment, and race differences in IQ: A commentary on Rushton and Jensen (2005). *Psychology, Public Policy, and Law, 11*, 302-310.

Nguyen, N. T. & McDaniel, M. A. (2003). Response instructions and racial differences in a situational judgment test. *Applied H. R. M. Research, 8*, 33-44.

Ogbu, J. U. (1978). *Minority education and caste: The American system in cross-cultural perspective*. New York: Academic Press.

Ogbu, J. U. (1994). From cultural differences to differences in cultural frames of reference. In P. M. Greenfield & R. R. Cocking (Eds.), *Cross-cultural roots of minority child development* (pp. 365-391). Hillsdale, NJ: Erlbaum.

Oswald, F. L., Schmitt, N., Kim, B. H., Ramsay, L. J., & Gillespie, M. A. (2004). Developing a biodata measure and situational judgment inventory as predictors of college student performance. *Journal of Applied Psychology*, *89*, 187-207.

Peeters, H. & Lievens, F. (2005). Situational judgment tests and their predictiveness of college students' success: The influence of faking. *Educational and Psychological Measurement, 65*, 70-89.

Pulakos, E. D. & Schmitt, N. (1996). An evaluation of two strategies for reducing adverse impact and their effects on criterion-related validity. *Human Performance, 9*, 241-258.

Rayder, N., Body, B., & Nimnicht, G. (1978). Assessing follow through: Changes in intelligence test scores over two and three years of experience in the responsive program. *Journal of Experimental Education, 47*, 60-66.

Raju, N. S. & Ellis, B. B. (2002). Differential item and test functioning. In F. Drasgow & N. Schmitt (Eds.) *Measuring and analyzing behavior in organizations: Advances in measurement and data analysis*. San Francisco: Josey-Bass.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: results and implications. *Journal of Educational Statistics, 4*, 207-30.

Robie, C., Zickar, M. J., & Schmit, M. J. (2001). Measurement equivalence between applicant and incumbent groups: An IRT analysis of personality scales. *Human Performance, 14*, 187-207.

Roth, P.L., Bevier, C.A., Bobko, P., Switzer, F.S., & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology, 54*, 297-330.

Roznowski, M. (1987). Use of tests manifesting sex differences as measures of intelligence: Implications for measurement bias. *Journal of Applied Psychology, 72*, 480-483.

Rushton, J. P. & Jensen, A. R. (2005). Thirty years of research on race differences in cognitive ability. *Psychology, Public Policy, and Law, 11*, 235-294.

Saad, S. & Sackett, P. R. (2002). Investigating differential prediction by gender in employment-oriented personality measures. *Journal of Applied Psychology, 87*, 667-674.

Scarr, S. (1994). Culture-fair and culture-free tests. In R. J. Sternberg (Ed.), *Encyclopedia of human intelligence*. New York: Macmillan.

Scheuneman, J. D. & Gerritz, K. (1990). Using differential item functioning procedures to explore sources of item difficulty and group performance characteristics. *Journal of Educational Measurement, 27*, 109-131.

Schmidt, F. L. (2002). The role of general cognitive ability and job performance: Why there cannot be a debate. *Human Performance, 15*, 187-210.

Schmitt, N. & Chan, D. (in press). Situational judgment tests: Method or construct? In J. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests*. Mahwah, NJ: Erlbaum.

Shivpuri, S. & Kim, B. (2004). Do employers and colleges see eye-to-eye?: College student development and assessment. *NACE Journal, 65*, 37-44.

Smith, K. C., & McDaniel, M. A. (1998). *Criterion and construct validity evidence for a situational judgment measure*. Paper presented at the 13th Annual Convention of the Society for Industrial and Organizational Psychology, Dallas, TX.

Society for Industrial and Organizational Psychology (2003). *Principles for the Validation and Use of Personnel Selection Procedures*. Bowling Green, Ohio: Author.

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.

Sternberg, R. J., Forsythe, G. B., Hedlund, J., Horvath, J. A., Wagner, R. K., Williams, W. M., Snook, S., & Grigorenko, E. L. (2000). *Practical intelligence in everyday life*. New York: Cambridge University Press.

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.

University of Illinois IRT Modeling Website (2005). Accessed May 15, 2005, at http://work.psych.uiuc.edu/irt

Wagner, R. K., & Sternberg, R. J. (1991). Tacit Knowledge Inventory for Managers. Unpublished research instrument available from authors.

Weekley, J. A. & Jones, C. (1999). Further studies of situational tests. *Personnel Psychology, 52*, 679-700.

Whitney, D. J. & Schmitt, N. (1997). Relationship between culture and responses to biodata employment items. *Journal of Applied Psychology, 82*, 113-129.

Zickar, M. J. (2002). Modeling data with polytomous item response theory. In F. Drasgow & N. Schmitt (Eds.) *Measuring and analyzing behavior in organizations: Advances in measurement and data analysis*. San Francisco: Jossey-Bass.

Table 1.

Twelve Content Areas of the Situational Judgment Test

Intellectual Behaviors

Knowledge, learning, and mastery of general principles

Gaining knowledge and mastering facts, ideas and theories and how they interrelate, and understanding the relevant contexts in which knowledge is developed and applied. Grades or GPA can indicate, but not guarantee, success on this dimension.

Continuous learning, and intellectual interest and curiosity

Being intellectually curious and interested in continuous learning. Actively seeking new ideas and new skills, both in core areas of study as well as in peripheral or novel areas.

Artistic and cultural appreciation

Appreciating art and culture, either at an expert level or simply at the level of one who is interested.

Interpersonal Behaviors

Appreciation for diversity

Showing openness, tolerance, and interest in a diversity of individuals (e.g., by culture, ethnicity, religion, or gender). Actively participating in, contributing to, and influencing a heterogenous environment.

Leadership

Demonstrating skills in a group, such as motivating others, coordinating groups and tasks, serving as a representative for the group, or otherwise performing a managing role

in a group.

Interpersonal skills

Communicating and dealing well with others, whether in informal social situations or more formal school-related situations. Being aware of the social dynamics of a situation and responding appropriately.

Social responsibility and citizenship

Being responsible to society and the community, and demonstrating good citizenship. Being actively involved in the events in one's surrounding community, which can be at the neighborhood, town/city, state, national, or college/university level. Activities may include volunteer work for the community, attending city council meetings, and voting.

Intrapersonal Behaviors

Physical and psychological health

Possessing the physical and psychological health required to engage actively in a scholastic environment. This would include participating in healthy behaviors, such as eating properly, exercising regularly, and maintaining healthy personal and academic relations with others, as well as avoiding unhealthy behaviors, such as alcohol/drug abuse, unprotected sex, and ineffective or counterproductive coping behaviors.

Career orientation

Having a clear sense of career one aspires to enter into, which may happen before entry into college, or at any time while in college. Establishing, prioritizing, and following a set of general and specific career-related goals.

Adaptability and life skills

Adapting to a changing environment (at school or home), dealing well with gradual or

sudden and expected or unexpected changes. Being effective in planning one's everyday activities and dealing with novel problems and challenges in life.

Perseverance

Committing oneself to goals and priorities set, regardless of the difficulties that stand in the way. Goals range from long-term goals (e.g., graduating from college) to short-term goals (e.g., showing up for class every day even when the class isn't interesting).

Ethics and integrity

Having a well-developed set of values, and behaving in ways consistent with those values. In everyday life, this probably means being honest, not cheating (on exams or in committed relationships), and having respect for others.

Note. These descriptions are based on those provided in Oswald et al. (2001) but have been revised slightly after surveying additional university administrators (Shivpuri & Kim, 2004).

Table 2

Location Parameters for Caucasians (C) and African Americans (AA) for Least, Most, and Composite Responses

	Least			Most			Composite		
Items	C	AA	Difference	C	AA	Difference	C	AA	Difference
1	.958	1.145	-.187	3.141	3.343	-.202	-.309	-.501	.192
2^{M}	-2.838	-1.422	-1.416				-1.717	-1.614	-.103
3	-.581	-.039	-.542	.304	.285	.019	-1.341	-1.326	-.015
4^{C}	-.201	-.370	.169	1.164	1.159	.005	-.620	-.655	.035
5^{C}	-.005	.006	-.011	1.758	1.347	.411	-.707	-.926	.219
6^{M}	-.188	-.510	.322				-1.232	-1.383	.151
7	3.824	3.364	.460	-.252	-.385	.133	-1.470	-1.476	.006
8^{C}	1.500	1.431	.069	2.122	1.901	.221	-1.123	-1.420	.297
9	.116	.456	-.340	1.293	-.304	**1.597***	-1.284	-1.558	.274
10	-1.084	-.166	-.918	.788	.221	.567	-.814	-1.126	.312
11	1.258	.257	**1.001***	1.428	1.084	.344	-.745	-1.055	.310
12	.484	.690	-.206	.575	.475	.100	-.576	-.352	-.224
13	.845	.137	**.708***	.615	.260	.355	-.732	-1.177	**.445***
14	2.420	1.713	.707	1.670	.279	1.391	.012	-.469	.481
15	.766	.871	-.105	-1.046	.209	**-1.255***	-1.986	-1.262	**-.724***

Note. * $p < .05$ indicates DIF. Superscripts indicate item misfit: C = composite, M = most, and L = least. A positive value for the difference in location parameters means the item favors African Americans and a negative value means the item favors Caucasians.

Table 2 (continued)

	Least			Most			Composite		
Items	C	AA	Difference	C	AA	Difference	C	AA	Difference
16	1.388	2.626	-1.238	.403	-.700	**1.103***	-1.855	-2.304	.449
17	.470	1.417	-.947	.179	-.034	.213	-.837	-.539	-.298
18[L]				1.404	1.262	.142	.290	.759	-.469
19	.380	.355	.025	-1.159	-.283	-.876	-2.238	-1.759	-.479
20	.243	-.536	**.779***	1.144	1.080	.064	-.864	-1.219	.355
22	-.904	-.463	-.441	-.832	-.777	-.055	-2.199	-1.447	-.752
23	2.827	2.890	-.063	2.852	2.754	.098	-.440	-.234	-.206
24	.039	-.688	**.727***	-.730	-.179	-.551	-1.874	-1.773	-.101
25	.837	.505	.332	.550	1.394	-.844	-.250	-.663	.413
26	.430	-.402	.832	-.732	-.575	-.157	-1.337	-1.201	-.136
27[C]	2.769	.894	**1.875***	2.231	1.877	.354	-.737	-1.302	.565
28	-.503	.123	-.626	-1.703	-.323	**-1.380***	-2.237	-1.571	**-.666***
29	-.665	.168	**-.833***	.315	.056	.259	-1.336	-1.729	.393
30[L]				-.512	-.490	-.022	-1.668	-1.682	.014
31[C]	.115	.358	-.243	2.475	2.483	-.008	-1.08	-1.103	.023
32	.034	-.679	.713	-.426	-.752	.326	-1.443	-1.666	.223
33[L]				.452	.212	.240	-.227	-.038	-.189
34	-.849	-1.024	.175	-.957	-.024	-.933	-1.647	-1.648	.001
35	-.491	-.746	.255	-.804	-.193	-.611	-1.447	-1.372	-.075
36	1.353	1.119	.234	1.185	.084	**1.101***	-.714	-1.120	.406

37	.302	-.239	.541	1.085	1.355	-.270	-.870	-1.334	.464
38	.866	1.084	-.218	.512	.770	-.258	-.679	-.310	-.369
39	-.369	.832	**-1.201***	-1.305	-.024	**-1.281***	-1.971	-.680	**-1.291***
Total DIF			7			6			4

Note. * $p < .05$ indicates DIF. Superscripts indicate item misfit: C = composite, M = most, and L = least. A positive value for the difference in location parameters means the item favors African Americans and a negative value means the item favors Caucasians.

Table 3

Items Showing DIF Within Each Content Area

		Least		Most		Composite	
Content	# of Items in Area	C	AA	C	AA	C	AA
Knowledge	6						
Continuous Learning	1						
Diversity	3				29		
Leadership	4		1,11				1
Interpersonal Skills	6		37	24	9	24	
Social Responsibility	3						
Health	2		4				
Career Orientation	5	32	15	25		25	
Adaptability	1						
Perseverance	6				31		
Ethics & Integrity	2	7		7		7	

Note. C = Caucasians favored, AA = African Americans favored. This table provides item numbers from Table 2 to show when DIF items across analyses were the same.

Appendix A.
Sample SJT Items

Knowledge
Your grade for a particular class is based on three exams, with no class attendance requirement. All of the homework requirements for the class are posted on the professor's web site. What would you do?

a. Attend class for as long as you feel that it is helping your grades.
b. Do all the homework but only go to some of the lectures. It's the exams that count.
c. Go to all the classes anyway. The professor may say something important.
d. Skip classes, but if you did poorly on the first exam, start going to classes.
e. There is no need to go to classes. Just get the homework done, and pass the exams.

Leadership
An important class project you have been working on with a group of other students is not developing as it should because of petty differences and the need of some members to satisfy their own agenda. How would you proceed?

a. Try to solve the group problems before starting on the work.
b. Work hard by yourself to make sure the project is finished, taking on others' share of the work if necessary.
c. Talk to the professor and get suggestions about solving the problem. If that doesn't work, try to switch groups or have an independent project.
d. Schedule a number of meetings, forcing the group to interact.
e. Take charge and delegate tasks to each person. Make them responsible for their part of the project.
f. Talk to the group and demand that they start working together.

Health
In the summer and fall, you walked to class and participated in various outdoor sports. When cold weather came, you took the bus and no longer participated in sports. You find that you are gaining weight. What action would you take?

a. Participate in indoor sports and start working out indoors.
b. Try not to eat as much or eat different kinds of food.
c. Walk to classes more, go to the gym and watch what you eat.
d. Work out in your room.
e. Talk to an expert in diets and see if you can find someone who will encourage you to start working out again.

What are you most likely to do?
What are you least likely to do?