This is to certify that the

thesis entitled

VIDEO-BASED VERSUS PAPER-AND-PENCIL METHOD OF ASSESSMENT
IN SITUATIONAL JUDGEMENT TESTS :
SUBGROUP DIFFERENCES IN PERFORMANCE AND EXAMINEE REACTIONS

presented by

DAVID CHAN

has been accepted towards fulfillment
of the requirements for

M.A. degree in PSYCHOLOGY

Neal Schmitt

Major professor

NEAL SCHMITT

Date MARCH 20, 1996

VIDEO-BASED VERSUS PAPER-AND-PENCIL METHOD OF ASSESSMENT IN SITUATIONAL JUDGEMENT TESTS : SUBGROUP DIFFERENCES IN PERFORMANCE AND EXAMINEE REACTIONS

By

David Chan

A THESIS

Submitted to

Michigan State University

in partial fulfillment of the requirements

for the degree of

MASTER OF ARTS

Department of Psychology

1996

ABSTRACT

VIDEO-BASED VERSUS PAPER-AND-PENCIL METHOD OF ASSESSMENT IN SITUATIONAL JUDGEMENT TESTS : SUBGROUP DIFFERENCES IN PERFORMANCE AND EXAMINEE REACTIONS

By

David Chan

Based on a conceptual distinction between test content and method of testing, the present study examined several theoretically and practically important effects relating race, reading comprehension, method of assessment, face validity perceptions, and performance on a situational judgement test using a sample of 241 psychology undergraduates (113 Blacks; 128 Whites). Results showed that the Black-White differences in situational judgement test performance and face validity reactions to the test were substantially smaller in the video-based method of testing than in the paper-and-pencil method. The Race X Method interaction effect on test performance was attributable to differences in reading comprehension and face validity reactions associated with race and method of testing. Implications of the findings were discussed in the context of research on adverse impact and examinee test reactions.

Dedicated to

Sapuan Bin Kasmari and Kong Kin Seng

ACKNOWLEDGEMENTS

This thesis could not have been successfully completed without the encouragement and support from a number of people. The most important individual is Neal Schmitt, the chair of my thesis committee. Neal has given me valuable guidance and support throughout the research process. He impresses me not just because of his professional expertise, but also because of his enormous efforts and great patience in the development of graduate students. The experience of working with Neal on the thesis and other research projects has made a great impact on my professional life. Neal has personified my ideal professor.

Dan Ilgen and Rick DeShon, the other two members on my thesis committee, have made very significant contributions to my graduate training. My first encounter with Dan was actually "on paper". He authored the I/O psychology textbook I used in my undergraduate days way back in Singapore. In fact, Neal and Dan were the primary reasons for my choice of graduate school. In retrospect, travelling thousands of miles across the ocean to Michigan State was well worth it. Dan's "motivational" seminar has resulted in lasting motivational effects on me. Rick was the first professor I worked with at Michigan State. His zeal for his work has always impressed me. From Neal and Rick, I have appreciated the meaning of "research interests".

Two other individuals, Kevin Ford and Steve Kozlowski, have indirectly contributed to the present piece of work. Both were instrumental in developing my fundamental expertise in I/O psychology without which I could not have completed the thesis so smoothly.

Finally, I must thank two groups of people who have provided me valuable social support. To fellow I/O graduate students at Michigan State, I am grateful for making me comfortable in a foreign land. To my dear friends in Singapore, I am grateful for the concern and constant update of things back home.

David Chan

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Overview

The present study examines the effects of a video-based versus a paper-and-pencil method of assessment on adverse impact and examinee reactions in a situational judgement test. The dependent variables of interest are test performance and examinee test reactions. Making the important distinction between test method and test content (Hunter & Hunter, 1984), test content is held constant across two different methods of testing so as to isolate subgroup differences on the dependent variables due solely to test methods.

The research problem leading to the present study will first be identified. The theoretical issues and practical concerns in personnel selection constituting the research problem will be explicated. Two conflicting goals in personnel selection will first be noted. The issue of adverse impact is then discussed and attempts to reduce adverse impact in selection is reviewed from the research on work samples and assessment centers. This will lead to the focal selection procedure in the present study namely, the situational judgement test which is becoming increasingly popular in the research and practice of personnel selection. The relationship between the logic of the test and its associated levels of adverse impact is discussed. The recent research on examinee test reactions will then be introduced. The frequently neglected but important issue of differential subgroup attitudes is examined and related to the important distinction between test content and the method of testing. Based on the literature review and

conceptual analysis in the Introduction, hypotheses for the present study are presented.

The hypotheses were tested in a sample of 241 undergraduates (113 Blacks, 128 Whites). Results supported the hypotheses. Limitations, contributions, and implications of the present study were discussed.

Two Conflicting Goals in Personnel Selection

A crucial element in the achievement of organizational goals is the selection of individuals with high ability to perform their jobs. Hence, the primary focus of personnel selection research and personnel selection procedures has always been the maximization of predictive efficiency by identifying and selecting individuals with the highest job-relevant ability. There has been a vast amount of empirical research on the validity and utility of selection procedures. Meta-analyses of these primary findings indicate that for a wide variety of jobs, valid measures of job-relevant ability dimensions can be developed and used to select high potential individuals. For example, paper-and-pencil measures of cognitive ability are valid predictors of most jobs in the US economy (Hunter & Hunter, 1984; Schmidt & Hunter, 1981). Assessment centers have consistently demonstrated validities for jobs involving managerial skills (Gaugler, Rosenthal, Thornton, & Bentson, 1987). Work samples (Schmitt, Gooding, Noe, & Kirsch, 1984) and biographical information (Reilly & Chao, 1982) are valid predictors of important job outcomes, and even interviews, when rigorously structured and administered, appear to be valid measures of job-relevant dimensions (McDaniel, Whetzel, Schmidt, & Maurer, 1994). Utility studies have also

shown that valid selection procedures can make substantial economic contributions to organizational productivity (e.g., Boudreau, 1983).

However, organizational productivity is not the only goal to be considered by the employer when selecting individuals. Schmitt and Noe (1986) noted that at least since the passage of the Civil Rights Act of 1964, political and legal demands have forced employers to consider a second and frequently conflicting goal namely, equal employment opportunities for various subgroups (minorities and women) in the American society. In 1965, President Johnson issued Executive Order 11246 which required all Federal contractors and subcontractors take affirmative action to ensure that employees are treated without regard to race, color, sex, religion, and national origin. This order, the passage of the Civil Rights Act of 1964, and subsequent court cases concerning charges of discriminatory use of tests constituted the zeitgeist for personnel researchers examining differences in validity of selection procedures across subgroups.

Subgroup Differences on Selection Tests

Schmitt and Noe (1986) provided a summary of the research and issues on subgroup differences in test performance and differences in validity of tests across subgroups including both differential validity and differential prediction. Whereas there is not much data regarding subgroup differences in validities of other predictors, the findings on subgroup differences (in particular, Black-White differences) in performance on paper-and-pencil measures of cognitive ability are well established.

There is little evidence of a Black-White difference in validity coefficients for paper-and-pencil measures of cognitive ability (i.e., little evidence of differential validity). Differential validity occurs when there is a significant difference between observed validities for two subgroups. Reviews of research have shown that differential validity is generally absent (Jensen, 1980; Linn, 1978) and when it is observed, the validity differences between Blacks and Whites are small and trivial (Cascio, 1982; Hunter, Schmidt, & Hunter, 1979). Moreover, Bobko and Bartlett (1978) have successfully argued that differential validity per se would not be a sufficient indicator of test bias. For example, different subgroup validity coefficients may result when two groups differ in variability even when their prediction systems are identical.

Although there is little evidence of differential validity, there is an extensive research demonstrating a sizable difference in test means of Black and White subgroups with Blacks on the average scoring about one standard deviation below Whites (e.g., Hunter & Hunter, 1984; Loehlin, Lindzey, & Spuhler, 1975; Schmidt, Greenthal, Hunter, Berner, & Seaton, 1977). Despite the absence of differences in subgroup validity coefficients, the use of paper-and-pencil measures of cognitive ability to select in a manner that optimizes predicted performance will still result in the hiring of a small number of Blacks relative to Whites because the Black subgroup mean score on the test is substantially lower than the White subgroup mean score. Hence, there is a conflict between the optimization of predicted performance (i.e., the goal of organizational productivity) and the goal of equal subgroup representation in selection.

A similar conflict is reached with respect to the assessment of differential prediction which is more directly related to issues of test bias than is differential validity. Differential prediction has now become the accepted way of evaluating test bias by most psychometricians. Evaluation of differential prediction involves the consideration of validity coefficients and standard errors of estimates and the regression line describing the predictor-criterion relationship. Predictions of performance are made using regression equations. According to the Cleary (1968) model of test bias, which is endorsed by both the Uniform Guidelines on Employee Selection (1978) and the Principles for the validation and use of personnel selection procedures (SIOP, 1987), a test is biased when a common regression equation results in either over- or under-prediction of subgroup performance, that is, a test is biased when there is differential prediction. Over-prediction for a protected minority group resulting from the use of a common regression line indicates test bias in the psychometric sense but is generally not considered a problem of fairness (SIOP, 1987). Hence, whereas test bias is a technical, psychometric issue, fairness is a social notion involving consideration of valued outcomes (SIOP, 1978).

The Cleary (1968) approach requires the use of separate subgroup regression equations when the equations are significantly different. The use of separate equations to provide a single rank order of applicants using predicted scores will result in hiring the best qualified individuals hence optimizing predicted performance but it will result in the selection of unequal proportions of members of various subgroups when subgroup mean performance differs. Schmitt and Noe (1986) noted that most research

evidence indicates that the use of a single common equation results in slight over-prediction of minority group performance whereas the use of separate equations results in average predicted performance for subgroups which is identical to the actual performance difference hence satisfying Cleary's (1968) criterion. Schmitt and Noe (1986) have also shown that the use of separate regression equations as prescribed by Cleary (1968) will result in selecting relatively few members of the lower scoring group (which is frequently the minority group) at all levels of selection ratios.

In short, paper-and-pencil measures of cognitive ability are valid predictors of job performance and generally unbiased toward minority subgroup members in the sense that their predicted performance matches their actual performance. However, sizable subgroup differences on test performance exist (with Blacks scoring on the average one standard deviation below Whites). Top-down selection on the basis of test scores results in the hiring of relatively small proportions of minority subgroup members. In most cases, the use of paper-and-pencil measures of cognitive ability in selection produces a high level of "adverse impact" on minority hiring rates (e.g., Hunter & Hunter, 1984; Schmidt et al. 1977) defined by the Uniform Guidelines on Employee Selection (1978) as the failure to meet the 4/5 rule, that is, the ratio of the proportion of minority applicants hired to majority applicants hired should not be lower than 4/5.

Attempts at Reducing Adverse Impact

The conflict between the goal of organizational productivity and the goal of equal subgroup representation prompted personnel researchers to try to develop valid predictors of performance that have levels of adverse impact lower than that associated with traditional paper-and-pencil measures of cognitive ability. A promising approach is the search for alternative predictor constructs. In this approach, researchers attempt to go beyond the construct of cognitive ability as assessed by traditional paper-and-pencil measures to measure other job-relevant abilities and attributes.

The logic for the construct-oriented approach to reducing adverse impact in selection is that paper-and-pencil measures of cognitive ability, while valid, may be measuring those determinants of job success on which subgroup differences are largest and conversely, they may fail to measure important determinants of job success on which such differences are smaller or nonexistent (Schmidt et al. 1977). However, the majority of the studies involving a search for alternative predictors have not adopted a construct-oriented approach. Instead, efforts have been directed at the development of alternative selection methods such as work samples, assessment centers, and biodata, and the efforts were often atheoretical. As argued later, this neglect of constructs resulted in a serious confound between method of testing and test content in assessment which has largely hindered our understanding of the nature of subgroup differences in performance on selection instruments. Many of these issues are best illustrated with the development and use of work sample tests as alternative predictors (to paper-and-pencil cognitive ability tests) of job performance.

The next section will summarize the research on the validity and adverse impact of work sample tests. The logic of work sample tests and the problems associated with their development and use will then be explicated. Assessment centers, an alternative predictor closely related to work samples will also be discussed to illustrate several issues and problems concerning the reduction of adverse impact. The discussion will lead to the consideration of situational judgement tests and their relationships to adverse impact and examinee test reactions which is the subject of the present study.

Work Sample Tests: Validity and Adverse Impact

In work sample tests, examinees are required to perform the same behaviors that they would be required to perform on the job. Several reviews have demonstrated that work sample tests can be at least as predictive of job performance as paper-and-pencil cognitive ability tests. Hunter and Hunter (1984) found that paper-and-pencil cognitive ability tests were about equally as valid as work sample tests. Schmitt et al.'s (1984) meta-analysis found that the validity of work samples were superior to those of biodata and cognitive ability tests. With respect to adverse impact, work samples appear to be advantageous compared to cognitive ability tests in that the mean difference between the scores of majority and minority subgroup members is typically less for work samples (Brugnoli, Campion, & Basen, 1979; Cascio & Phillips, 1979; Schmidt et al. 1977; Schmitt, Clause, & Pulakos, 1996; Wigdor & Green, 1991). For example, Schmidt et al. (1977) compared the adverse impact of a content-valid work

sample test of metal trade skills to that of a well-constructed content-valid paper-and-pencil achievement test for the same technical area. They found the typical one standard deviation Black-White subgroup difference for the paper-and-pencil test but found no significant difference between Blacks and Whites for the work sample test. Bernardin's (1984) meta-analytic review of Black-White differences on work sample tests found an average difference of .54 standard deviation units favoring Whites.

The Logic and Problems in Development and Use of Work Samples

In order to explain the positive results of the work sample test regarding its validity and small adverse impact, one needs to examine the logic of the development and use of work samples. The interest in work samples could in part be traced to Wernimont and Campbell (1968) who contended that samples of the kinds of behaviors actually required to be performed on the job would predict future job performance better than scores on typical cognitive ability tests. The authors argued that scores on ability tests are merely "signs" which are less similar to and hence less related to actual job performance compared to "samples" of the work on the job. The implicit assumption is that the more similar a test is to the actual job, the higher the validity of the test. In accounting for the predictive success of work sample tests, Asher and Sciarrio (1974) stated that a strong relationship between the content of the job and the content of the selection method must exist for high predictive validity to occur. Smith and George (1992) argued that Asher and Sciarrino's (1974) "point to point " validation theory on work samples can be used as an explanation for the

success and failure of most selection methods.

However, the notion of "similarity" between test and actual job has never been sufficiently explicated in most studies which examined tests purportedly similar to actual job content. This is certainly true in the case of work sample tests. Given that performance on most jobs is multidimensional, a work sample successfully replicating a portion of the job is almost always multidimensional. However, little if any work has been directed to understanding the nature of the constructs measured in work sample tests.

Schmitt et al. (1996) reviewed studies on subgroup differences published from 1964 to 1994 in three major journals concerned with personnel selection (Journal of Applied Psychology, Journal of Occupational and Organizational Psychology, and Personnel Selection) and attempted to ascertain the nature of the constructs measured and methods used in those studies. With regard to work samples, the authors found that it was almost never clear what construct(s) were measured. In the same review, the authors also noted that the data available regarding the lower adverse impact associated with work samples relative to paper-and-pencil cognitive ability tests are not very useful in providing us an understanding of the reasons for the reduction in adverse impact. This is due to an inherent confound between method and test content in almost all studies comparing subgroup differences on the two types on tests. In these studies, work samples and cognitive ability tests differed in the method of testing (e.g., paper-and-pencil versus actual task performance) and presumably, the nature of the constructs measured (e.g., general cognitive ability versus interpersonal-oriented

dimensions) due to different item content between the two tests.

The distinction between method and content (Hunter & Hunter, 1984) is crucial to the study of reduction in adverse impact. If method and test content is disconfounded in a study, then subgroup differences due to method and subgroup differences due to test content can be isolated. In principle, we can then reduce or even eliminate adverse impact by changing method of testing or test content depending on the job-relevance of the given constructs. For example, two different methods of testing may have the same test content measuring the same job-relevant construct but one method produces less adverse impact than the other. Adverse impact due solely to method of testing can then be eliminated by using the method with lower adverse impact assuming that method is job-irrelevant.

On the other hand, by controlling method, we may be able to ascertain different test contents that differ in the size of subgroup differences they produce. For example, subgroup differences may be smaller for test content tapping interpersonal skills than test content tapping cognitive constructs (Hough, 1994; Hough, Eaton, Dunnette, Kamp, & McCloy, 1990). Assuming both types of constructs are job-relevant, adverse impact can be reduced and validity can be increased by expanding the predictor space beyond the measurement of cognitive constructs to include the measurement of interpersonal skills constructs.

The present study differentiates method from content by comparing two different testing methods (paper-and-pencil versus video-based assessment) with the same set of test items. The importance of the method-content distinction in the

present study will be elaborated later. In short, in terms of our understanding of the smaller adverse impact associated with work samples relative to cognitive ability tests, more work is certainly needed on the nature of the constructs measured in work samples, their representativeness of the job, and issues relating to the physical fidelity and psychological fidelity of the simulation (McHenry & Schmitt, 1994).

There are several practical problems that have limited the use of work samples. Despite its validity and low adverse impact, many organizations have not incorporated work samples into their selection procedures due to the high cost of testing. Work samples are often expensive to develop and administer, especially when raters are required. Many work sample tests are administered one on one by a test administrator who often has to score the results by hand (McHenry & Schmitt, 1994). To ensure reliability, more raters are required which increases the cost of testing. Costs are further increased when complex administration and scoring procedures demand rigorous assessor training (Wigdor & Green, 1991). In certain cases, work sample tests may not be practical due to the potential danger to the applicant inherent in the tasks. Jobs involving high physical demands may be least practical for the development of work sample tests and yet these may be the jobs where work samples are most predictive. Finally, some jobs may be sufficiently technical and involve a substantial amount of job-specific knowledge such that it would not be possible to develop a work sample that is representative of a significant portion of the job and at the same time applicable to applicants (who do not have the knowledge and skills of the experienced incumbents).

Assessment Centers

Several issues and problems concerning the reduction of adverse impact using alternative predictors can be illustrated with the development and use of a selection instrument closely related to work samples namely, the assessment center. Although primary research and reviews in personnel selection have almost always treated assessment centers as a type of predictor distinct from work samples, the two predictors have much in common. Both assessment centers and work samples are based on a behavioral sampling assumption and they share the basic tenet of the behavioral consistency approach that the best predictor of future performance is present or past performance or behavior of the same type. Both are simulations in the sense that the task stimuli are constructed such that they mimic actual job situations and elicit responses which are purported indicators of how assessees would handle the task situations if they were actually occurring on the job. Assessment centers are more like "samples" than like "signs" in the sense distinguished by Wernimont and Campbell (1968). Both work samples and assessment centers are almost always multidimensional reflecting the multidimensionality of the target job they mimic. Both also tend to have high face validity. Both often require trained raters who are also subject matter experts on the target job and both are expensive to develop and administer. The distinguishing feature of the assessment center is its multiexercise-multirater methodology. Also, although in principle the multiexercise-multirater methodology can be applied to almost any job, assessment centers have been historically restricted to the assessment of general managerial dimensions. Because it

typically assesses general managerial dimensions as opposed to some job-specific technical knowledge and skills, the assessment center is less likely to have the problem of inapplicability to inexperienced applicants faced by many work samples alluded to earlier.

With respect to validity and adverse impact, research on assessment centers has demonstrated a pattern of findings similar to that of work samples. Like work samples, validities obtained for assessment centers are at least comparable to those observed for cognitive ability tests. At least two meta-analyses have found substantial validities for assessment centers. Schmitt et al. (1984) found an average validity of .41 across 21 studies. Gaugler et al. (1987) obtained an average validity of .34 based on 107 validity coefficients for various performance criteria from 50 studies.

Like work samples, typical Black-White subgroup differences in assessment center performance are also substantially smaller than the one standard deviation difference observed for cognitive ability tests (e.g., Huck & Bray, 1976). Based on a sample of 2,910 candidates who were assessed for school administrator positions in 25 different assessment centers using the same set of exercises and dimensions, Schmitt (1993) found significant mean differences between Black and White subgroup members for 10 of 13 dimensions ranging between two-thirds to three-fourths of a standard deviation in favor of Whites. In short, assessment centers by no means eliminate adverse impact but they tend to have Black-White differences substantially smaller than those for cognitive ability tests.

Whereas studies of work samples have tended to neglect the issue of

constructs, a substantial amount of research has been devoted to the study of construct validity of the dimensional ratings in assessment centers. However, our understanding of the nature of the constructs tapped in assessment centers is no better than the case in work samples. Multitrait-multimethod studies have consistently reported low construct validity of dimensional ratings and factor analyses of these ratings produced "exercise factors" rather than dimensional factors (e.g., Chan, in press; Sackett & Dreher, 1982; Sackett & Harris, 1983; Schneider & Schmitt, 1992; Turnage & Muchinsky, 1982). In describing the lack of construct validity in assessment center research, Klimoski & Brickner (1987) noted that we know assessment centers work in the sense that they have predictive validity but we do not know why insofar as we have little understanding of the nature of the constructs tapped by assessor ratings.

Just as in the case of work samples, it is tempting to attribute the smaller subgroup difference observed in assessment center performance (relative to paper-and-pencil cognitive ability tests) to the nature of the constructs tapped by the test content. Like work samples, one may hypothesize that the multidimensionality of assessment centers included both cognitive and non-cognitive constructs (e.g., interpersonal dimensions) and that subgroup differences on non-cognitive constructs may be smaller or even non-existent compared to cognitive constructs such that the overall ratings in assessment centers exhibit lower adverse impact relative to paper-and-pencil measures of cognitive constructs. However, as mentioned earlier, the test of such a hypothesis would require a design eliminating the method-content confound in the comparison between assessment center performance and performance on cognitive ability tests.

Unfortunately, a fully-crossed content by method factorial design is often not feasible. For example, in a paper-and-pencil methodology, it is difficult to develop test content tapping many of the usual assessment center dimensions (e.g., leadership, decisiveness) and sometimes impossible to do so (e.g., oral communication).

Schmitt et al. (1996) reviewed studies on subgroup differences and found no study which employed the method by content design. However, they did find one unpublished study (Goldstein, Braverman, & Chung, 1993) reporting subgroup differences measured using different methods. The Goldstein et al. (1993) study will now be described in order to discuss the core issues associated with the method by content design approach to examining subgroup difference. Some of the problems with the design used in Goldstein et al. (1993) will be addressed in the present study.

The Goldstein et al. (1993) study

The purpose of Goldstein et al. (1993) was to examine the effects of different testing methods on subgroup differences. The authors attempted to address the "method versus content" issue by developing four tests that purportedly assess the same six abilities. The sample consisted of 29 Whites and 13 Blacks who were being assessed for promotion in a police organization. The four tests used, which were construed as work samples by the authors, were similar to the typical exercises in an assessment center. They were a written in-basket test, a role-play exercise in which the examinee conducts a performance appraisal counseling session, a simulation planning exercise requiring the examinee to develop contingency plans to a

hypothetical event, and a simulation exercise in which the examinee supervises activities associated with the event that he or she had prepared in the simulation planning exercise. The six abilities assessed across all four tests were the ability to pay attention to details, to adjust communication to level of understanding of other person, to communicate using proper grammar and wording, to put materials in a logical sequence, to adjust action or decision in light of new information, and to maintain composure in stressful situations.

Citing Helms (1992), Goldstein et al. argued that the African-centered values and beliefs of Blacks emphasize communalism, movement, and orality which would in turn influence their test-taking performance. Accordingly, Blacks have a disadvantage on paper-and-pencil tests compared to Whites due to the strong written component requirement for successful performance on such measures. The written component is construed as a requirement of the test method and is not part of the construct intended to be assessed by the test content.

The authors hypothesized that a testing method requiring a written response mode favors Whites over Blacks whereas tests that were more interactive, behaviorally-oriented, and aurally-/orally-oriented would exhibit less adverse impact. Hence, it was predicted that the written in-basket test would have a higher level of adverse impact relative to the other three tests which were more interactive, behavioral, and aural/oral in nature. The results were consistent with the hypothesis. The written in-basket test had a substantially higher level of adverse impact (.47 to .87, average = .65) when compared to the simulation planning exercise (.41 to .64,

average = .48) and the simulation exercise (.22 to .36, average = .30). For the role play which is presumably the most interactive-oriented exercise, Blacks performed better than Whites (.38 to .64, average = .58).

Schmitt et al. (1996) noted several limitations with Goldstein et al.'s (1993) study. The sample sizes were small with only 13 Blacks and 29 Whites. No reliability estimates were reported for the various measures. With low reliabilities, true subgroup differences will not be detected. It is possible that some of the more interactive measures (e.g., simulation exercise) are substantially less reliable than paper-and-pencil measures such that true subgroup differences were not detected on the former. That is, it was not clear if Goldstein et al.'s (1993) findings were due to true subgroup differences or simply an artifact of differential reliability in measurement. In the present study which compared two methods of assessment in a situational judgement test, the reliabilities of each measurement method were estimated so that effect sizes could be corrected for unreliability in measurement. Adequate sample sizes were also employed to ensure sufficient power.

Schmitt et al. (1996) also noted that there was no evidence establishing the equivalence of constructs across methods in Goldstein et al's. (1993) study. This is an important concern because, as argued earlier, the adequacy of a method by content design for the isolation of method sources and content sources of subgroup differences presupposes an equivalence of constructs across methods when test content is held constant across methods. In Goldstein et al. (1993), the content of the task stimuli (i.e., test content) appeared to be quite different across test methods. For example, it

was not clear if the "ability to maintain composure under stressful situations" elicited by the preparation of memos in the in-basket test (and rated by assessors) was in fact the same construct as the purportedly same dimension elicited (and rated) by the interactions in the counseling situation of the role play exercise. In the present study on situational judgement, the issue of construct equivalence was addressed by administering the same test items using two different methods of stimulus presentation and empirically testing factorial invariance of test responses across the two methods.

Another limitation of Goldstein et al. (1993) was that the ability dimensions described were relatively specific and their results may not be generalizable to the broader psychological constructs of interest typically assessed by the common predictor instruments in personnel selection such as cognitive ability tests, personality measures, and work samples. The present study on situational judgement employed more global constructs such as interpersonal skill dimensions of conflict resolution and empathy.

In order to provide a more rigorous test of the hypothesis that a significant amount of the Black-White difference in performance on paper-and-pencil tests is due solely to the reading/written requirements inherent in the method of testing and independent of the construct measured, the present study also administered a reading comprehension test to both Blacks and Whites. The hypothesis would predict lower reading comprehension scores for Blacks and that the Black-White subgroup difference in performance on the paper-and-pencil method of testing will be reduced when reading comprehension is controlled.

The Present Study

The present study examined the effects of a video-based versus a paper-and-pencil method of assessment on adverse impact and examinee test reactions in a situational judgement test. With respect to adverse impact, test content (and presumably, the constructs measured) was held constant across two different methods of testing so as to isolate subgroup differences due solely to test methods. As mentioned earlier, construct equivalence across methods was empirically tested. Reliabilities of measurement were estimated to obtain corrected effect size estimates. A reading comprehension test was administered to provide an additional test of the hypothesis that a significant amount of the Black-White difference in performance on paper-and-pencil tests is due solely to the reading/written requirements inherent in the method of testing independent of the test content. As discussed thus far, the study addressed the issues and problems associated with evaluating the effect of test method and test content on the size of subgroup differences in test performance. The use of the present situational judgement test circumvented many of the conceptual and practical problems associated with typical work samples and assessment centers explicated earlier. The logic and research on situational judgement tests and their relationship to adverse impact will be discussed next. Examinee test reactions, the second dependent variable in the present study, will then be introduced. The study of test reactions has become increasingly important in recent personnel selection research and the links between test reactions and the method-content distinction will be explicated.

The Logic of Situational Judgement Tests

In a typical situational judgement test, examinees are presented with a hypothetical scenario describing a work situation in which a problem has arisen. The work situation may be a possible actual situation on the target job or a situation constructed such that it is psychologically isomorphic to an actual situation. The latter would address the problem faced by typical work samples concerning inapplicability of test items to inexperienced applicants due to the requirement of job-specific knowledge and experience on some jobs. Either way, the work situations on the test are developed on the basis of job analysis data often including a critical-incident analysis involving subject matter experts. The individual situational judgement problem is almost always multidimensional in nature in the sense that an adequate solution or handling of the problem would involve several ability and skill dimensions.

Alternative responses are presented to the examinee following the description of the situation. Examinees' scores on the test are computed based on their endorsement of the responses. In tests employing a forced-choice format, examinees are typically asked to choose the most effective response, or to choose the most effective response and the least effective response. In another format (the format used in the present study), examinees are asked to rate each response in terms of its effectiveness usually using some form of a Likert-type scale. The scoring key is developed from prior effectiveness ratings of response alternatives obtained from subject matter experts. The decision rules for identifying the most or least effective response or arriving at the score for each effectiveness rating given by examinees vary

from test to test. Regardless of the precise rules used, statistical analyses and sometimes content analyses are performed on the subject matter expert ratings to ensure reliability and agreement in the ratings used for the development of the scoring key.

Often, the objective of developing a situational test is to sample behaviors from the domain of job performance rather than measuring any particular construct or predispositional sign. Hence, like work samples, situational judgement tests are more like "samples" than like "signs". However, Motowidlo, Dunnette, & Carter (1990) noted that it would be interesting to discover what constructs are measured by the test. The importance of construct-orientation and the distinction between method and content for the examination of adverse impact has been discussed earlier. Identifying the nature of the constructs measured in situational judgement tests will provide a better understanding of the causes of adverse impact and help in the development of ways of reducing the level of adverse impact associated with some given selection instrument.

Situational Judgement Tests: Simulation Fidelity and Predictive Validity

Work samples, assessment centers, and situational judgement tests may all be construed as forms of simulations. In these simulations, task stimuli are constructed such that they mimic actual job situations and elicit responses which are purported indicators of how assessees would handle the task situations if they were actually occurring on the job (Motowidlo et al. 1990). Work samples are on the high end of

the continuum of simulation fidelity because they use very realistic materials to represent the task situation and examinees may respond in a manner almost identical to the way they would if they were actually on the job.

As tests move toward the low end of the fidelity continuum, stimuli and responses are less faithful approximations of actual job stimuli and responses. The situational interview (Latham & Saari, 1984; Latham, Saari, Pursell & Campion, 1980; Weekley & Gier, 1987) is a well-known example of a simulation on the lower end of the fidelity continuum. Latham et al. (1980) reported a situational interview with a validity of .46 and Latham & Saari (1984) reported a validity of .14.

Motowidlo et al. (1990) developed a paper-and-pencil type of situational judgement test which they termed a "low-fidelity" simulation. In this test, the task stimulus (i.e., the work situation) is presented in a written form and examinees are required to endorse alternative responses described also in written form. The test resembles similar situational inventories developed in early research such as the Supervisory Practices Test (Bruce & Learner, 1958), the "How supervise?" (File & Remmer, 1971), and the Leadership Evaluation and Development Scale (Tenopyr, 1969). The paper-and-pencil method of administering the situational judgement test in the present study is a type of "low-fidelity " simulation with a format similar to the test developed by Motowidlo et al. (1990) except that instead of a forced-choice response format, the present test requires examinees to give effectiveness ratings for each of the alternative responses.

Motowidlo et al. (1990) noted that although simulations with higher fidelity

should be better predictors of actual job performance than those with lower fidelity according to the basic tenet of behavioral consistency, there have been no systematic studies of the relationship between differences in fidelity and incremental predictive value. Such high fidelity simulations as work samples and assessment centers are expensive to develop and administer and the cost of developing such simulations may not offset the gain in predictive value over lower fidelity simulations (Motowidlo et al., 1990).

Whereas it is expensive and often not feasible to administer work samples or assessment centers to a large group of examinees in one testing session, situational judgement tests can be administered to relatively large numbers of examinees in one session. In the case of a paper-and-pencil format of the test, the scale of testing effort and expense is identical to traditional paper-and-pencil measures of cognitive ability tests or personality tests. Moreover, work samples and assessment centers almost always require substantial involvement of subject matter experts for rating or scoring of individual examinee performance at the time of testing and ongoing assessor training costs can be high. On the other hand, the primary involvement of subject matter experts in the situational judgement test is in the development of the test stimulus (work situations) and scoring key. Hence, from a practical viewpoint, it is worthwhile to explore the predictive validity of low fidelity simulations such as the situational judgement test.

Using a sample of approximately 120 management incumbents, Motowidlo et al. (1990) found positive validities for their low fidelity situational judgement test in

predicting supervisory ratings of performance (.28 to .37, $p < .01$). Further evidence of validity for the test were provided in Motowidlo & Tippins (1993) in which two studies were reported. Study 1 employed a predictive validation design and found an average validity of .25 predicting supervisory performance ratings in a sample of 36 management applicants. Study 2 employed a concurrent validation design and found an average validity of .20 predicting supervisory performance ratings in a sample of 109 to 128 marketing incumbents. Pulakos, Schmitt, & Keenan (1994) developed a situational judgement test similar in format to Motowidlo et al.'s (1990) low-fidelity simulation test. Using a sample of incumbents from a large federal investigative agency, they found significant validities for the test in predicting two performance criteria namely, core investigative proficiency (.20) and effort and professionalism (.13).

Situational Judgement Tests and Adverse Impact

Motowidlo et al. (1990) found a Black-White difference of .21 standard deviation favoring Whites in their sample of incumbents and a difference of .38 standard deviation favoring Whites in their sample of applicants. Although these differences were nonsignificant, there is a caution against concluding that situational judgement tests successfully eliminated adverse impact. The number of Blacks in Motowidlo et al.'s (1990) samples were small (ranging from 21 to 31) and the power to detect a difference of .5 standard deviation was only between 47% and 68% (Cohen, 1977). Of the two studies reported in Motowidlo & Tippins (1993), one

provided no information on Black-White differences as the sample of Blacks was too small for subgroup analysis (N = 16). The other study reported that Blacks scored lower than Whites by .38 standard deviation (44 Blacks vs 178 Whites). Weighting the Black-White differences reported in Motowidlo et al. (1990) and Motowidlo & Tippins (1993) by their sample sizes yielded an average adverse impact of .32 standard deviation (total of 97 Blacks vs 378 Whites). In the situational judgement test developed by Pulakos, Schmiitt, & Keenan (1994), Blacks scored lower than Whites by .41 standard deviation (100 Blacks vs 259 Whites).

Video-Based Situational Judgement Test

The above review showed that adverse impact levels of the paper-and-pencil type of situational judgement test appear to be substantially lower than the typical one standard deviation for cognitive ability tests but the size of the Black-White difference is still considered at least moderate and is practically significant. A primary purpose of this study was to examine the possibility of reducing the Black-White difference on the situational judgement test by simply changing the method of stimulus presentation from the paper-and-pencil delivery to a video-based delivery while keeping test content constant. The theoretical rationale for this hypothesis has been explicated earlier in the discussion of the Goldstein et al. (1993) study. By replacing the paper-and-pencil method which requires reading comprehension with the more interactive, behavioral, and orally-/aurally-oriented video-based method, the Black-White difference in test performance should be reduced.

Although the advantages in use of video-based testing in personnel selection have been alluded to as early as in Thorndike (1949), its actual use is relatively new and there is an insufficient research base evaluating the psychometric properties and adverse impact of the assessment method. However, the few studies conducted did report some encouraging results for a video-based method of presenting the situational judgement test. Based on a KSAO analysis of 50 customer service jobs, Wilson Learning (1990) developed a video-based situational judgement test for the assessment of customer service skills. Using performance ratings as the criterion, the test was found to have a validity of .40 for a sample of 126 Canadian employees and .34 for a sample of 60 American employees. In another video-based test developed for transit operator selection, Smiderle, Perry, & Cronshaw (1994) reported a significant negative validity using number of complaints as the criterion but no significant correlations were found between test scores and two other criteria namely, commendations and a performance composite. Dalessio (1994) also found a significant average validity of .17 for a video-based test predicting turnover a year later using several samples of insurance agents (total $N = 677$).

The present author located only one published study reporting the adverse impact level of the video-based situational judgement test. Smiderle et al. (1994) found no significant Black-White difference in test performance (46 Blacks vs. 267 Whites). However, the result was not corrected for unreliability of measurement. The low reliability of the test (alpha = .47) certainly attenuated the true Black-White difference. Moreover, the present author performed a power analysis (Cohen, 1988)

on the data and found that the study had only a power of approximately 59% to detect a moderate effect size ($\underline{d}$ = .5) at α = .05. Hence, more research is needed to ascertain the adverse impact level of video-based situational judgement tests. The present study examined Black-White differences in performance on a video-based assessment and compared it with the difference on a paper-and-pencil format of the same situational judgement test. A priori power analyses were conducted to ensure adequate sample sizes and reliabilities of the two measurements were estimated to correct for attenuation due to unreliability.

The present study developed two formats of a single situational judgement test, differing in the method of testing (video-based versus paper-and-pencil presentation of the work situations) with test content held constant. As discussed earlier, Helms (1992) theorized that African-centered values and beliefs of Blacks emphasize communalism, movement, and orality at the expense of reading comprehension. The lack of emphasis on reading comprehension in turn influences their test-taking performance resulting in Blacks having a disadvantage on paper-and-pencil tests compared to Whites due to the strong written component requirement for successful performance on such measures. Reviews have cumulated an extensive research evidence showing a significant and substantial Black-White difference on paper-and-pencil measures of cognitive-oriented constructs in favor of Whites, that is, a high level of adverse impact exists. Results from Motowidlo et al. (1990), Motowidlo & Tippins (1993), and Pulakos et al. (1994) indicated that Blacks also score lower than Whites on a paper-and-pencil type of situational judgement test. Prior to testing the

primary hypotheses concerning effects of test method on adverse impact, it was necessary in the present study to first replicate the previous findings that Blacks perform significantly poorer than Whites on a situational judgement test presented in a paper-and-pencil format.

Goldstein et al. (1993) and Schmitt et al. (1996) have argued that a testing method loaded with a strong reading/written component would tend to favor Whites over Blacks whereas tests that were more interactive, behaviorally-oriented, and aurally-/orally-oriented would exhibit less adverse impact. Based on this argument and Helm's (1992) theory, it was predicted that for performance on the situational judgement test,

H1: A Race X Method interaction effect on test performance will occur. Specifically, the Black-White difference in test performance using the video-based method will be smaller than that using the paper-and-pencil method.

The nature of the expected interaction is depicted in Figure 1.

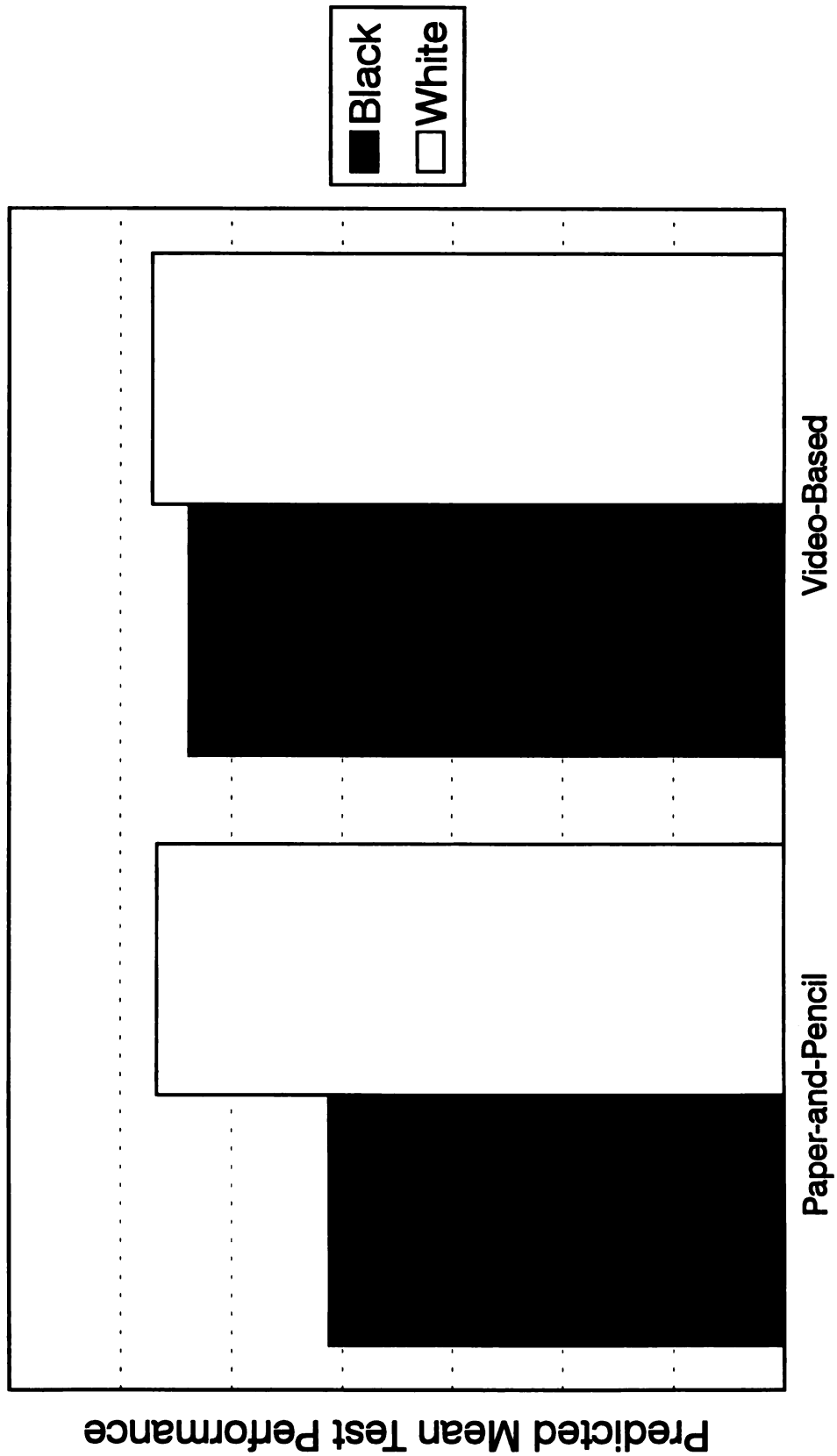

Figure 1. Hypothesis 1: Predicted Race X Method Interaction on Situational Judgement Test Performance.

It was argued earlier in the paper that a significant amount of the Black-White difference in performance on paper-and-pencil tests could be due solely to the reading comprehension inherent in the method of testing independent of the test content. Two hypotheses were derived from this argument. One hypothesis related performance on the test to the method of testing and individuals' reading comprehension ability whereas the other hypothesis related test performance to method of testing, reading comprehension, and racial subgroup membership. With respect to method of testing and reading comprehension ability, it was expected that an individual's performance on the situational judgement test would be affected by his or her reading comprehension ability when the test was administered using the paper-and-pencil method but no such effect would exist when the test was administered using the video-based method. Hence, it was predicted that for performance on the situational judgement test,

H2: A Method X Reading Comprehension interaction effect on test performance will occur. Specifically, test performance will be positively and significantly correlated with reading comprehension ability in the paper-and-pencil method of testing; whereas there will be no significant correlation between test performance and reading comprehension ability in the video-based method of testing.

The nature of the expected interaction is depicted in Figure 2.

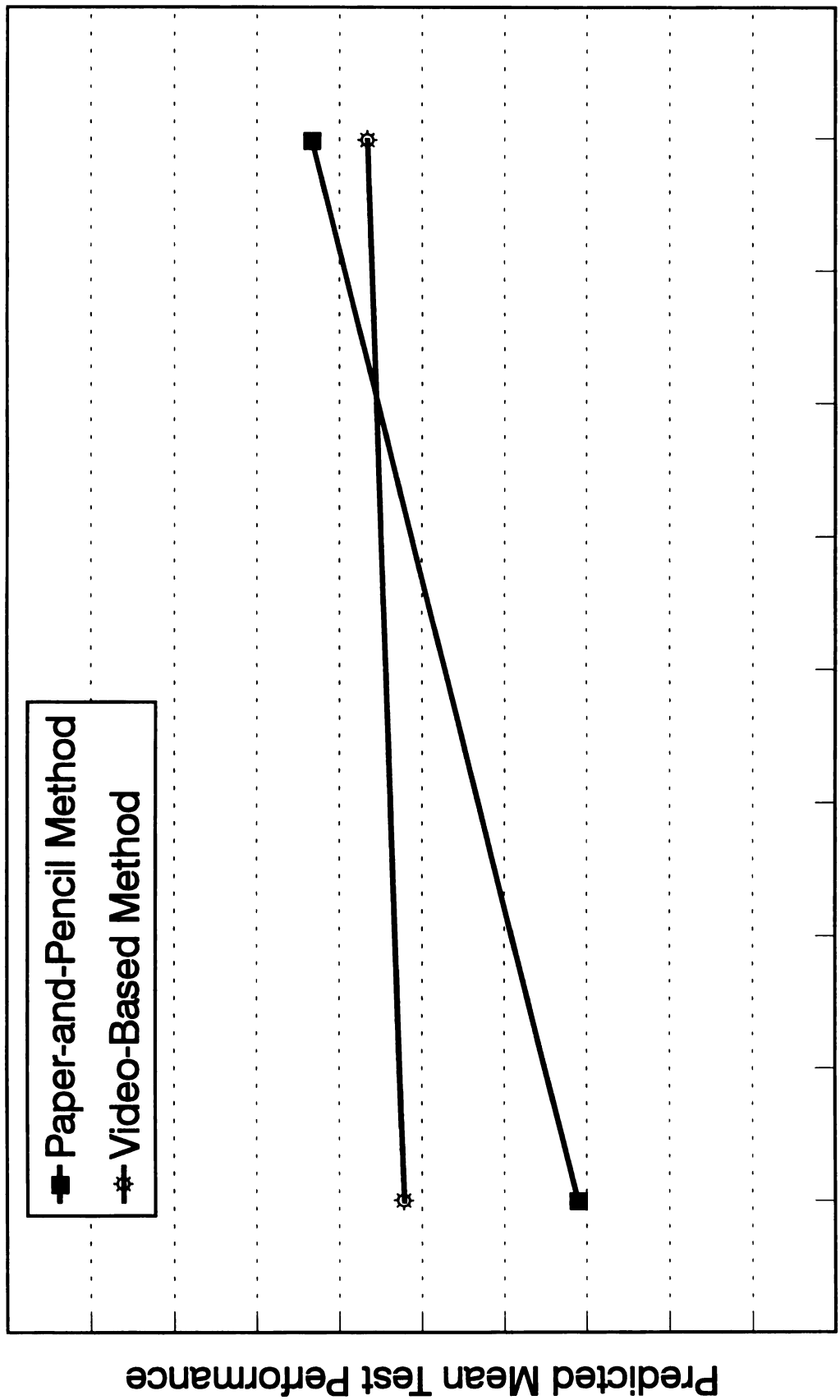

Figure 2. Hypothesis 2: Predicted Method X Reading Comprehension Interaction on Situational Judgement Test Performance.

Previous research has shown that a Black-White subgroup difference exists on reading comprehension tests favoring Whites (e.g., Matthews, 1991; Scott, 1987). A Black-White difference in reading comprehension scores favoring Whites is expected to be replicated in the present study. Thus, a significant amount of the Black-White difference in performance on paper-and-pencil tests could be due solely to the reading comprehension requirements inherent in the method of testing independent of the test content. That is, a substantial amount of the Race X Method interaction effect on test performance hypothesized in H1 could be due solely to the Method X Reading Comprehension interaction effect on test performance hypothesized in H2. Hence, it was predicted that:

H3: The Race X Method interaction effect would diminish after controlling for the Method X Reading Comprehension interaction effect on test performance.

The predicted reationship between Race X Method interaction and Method X Reading Comprehension interaction on test performance is depicted in Figure 3.

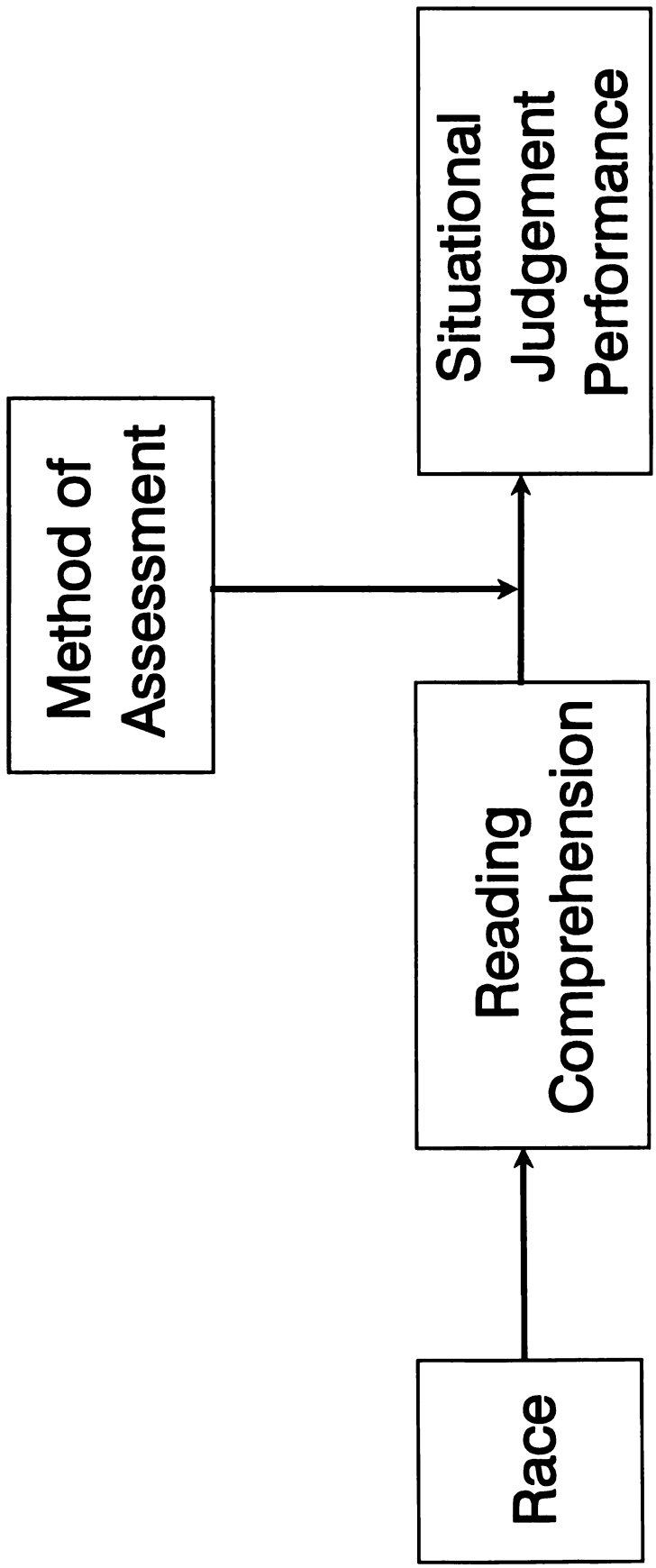

Figure 3. Hypothesis 3: Method X Reading Comprehension Interaction as an Explanation for Race X Method Interaction on Situational Test Performance.

Examinee Test Reactions

Research on test validity and adverse impact has tended to examine predictor adequacy and fairness from the organizational and psychometric perspectives. Recent research in personnel selection has begun to focus more attention on applicant reactions or examinee attitudes to selection procedures (e.g., Arvey, Strickland, Drauden, & Martin, 1990; Gilliland, 1993; Gilliland, 1994; Macan, Avedon, Paese, & Smith, 1994; Schmitt & Gilliland, 1992; Schmitt, Gilliland, Landis, & Devine, 1993). This notion of perceived test adequacy and fairness has been termed "social validity" (Schuler, 1993), "impact validity" (Iles & Robertson, 1989) and the "social side of selection" (Herriot, 1989).

Examinee reactions to selection procedures could be organizationally relevant in that it could affect applicant and employee behaviors (Arvey & Sackett, 1993; Gilliland, 1994). Premack and Wanous (1985) and Robertson and Smith (1989) argued that assessment situations serve as a preview of the organization and Schuler and Fruhner (1993) noted that selection instruments can be used as instruments for personnel marketing. Smither, Reilly, Millsap, Pearlman, and Stoffey (1993) outlined three possible practical effects of applicant reactions. First, reactions can indirectly influence applicant pursuit or acceptance of job offers through organizational attractiveness. Second, reactions may relate to the likelihood of litigation and the success of the defense of the selection procedure. Third, reactions may indirectly affect both validity and utility through motivation in test performance and loss of qualified applicants respectively. In short, examinee test reactions are of interest

because they constitute a critical component of the recruitment-selection process.

Face Validity and Predictive Validity Perceptions

Whereas there has been extensive research on the validity of work samples and assessment centers, there are few studies which systematically investigate examinee reactions to these predictors such as face validity and predictive validity perceptions. Schmitt and Gilliland (1992), Gilliland (1993), and Gilliland (1994) have attempted to relate organizational theories of distributive and procedural justice to examinee test reactions and Schuler (1993) has proposed a model of social validity. However, in general the research on examinee reactions to selection procedures has been fragmented and atheoretical. Most research has been focused on the description of examinee attitudes or reactions to different selection tests and compared reactions across tests. There is a need to integrate studies of examinee attitudes into the broader selection framework. In the present study, the investigation of examinee test reactions is integrated into the selection framework by analyzing attitudes by race and examining relationships between attitudes and adverse impact and method of testing.

Research on examinee reactions has focused on face validity and little effort has been directed to investigating perceived predictive validity. Face validity (also known as perceived content validity) refers to the extent to which examinees perceived the content of the selection procedure to be related to the content of the job. Perceived predictive validity refers to the extent to which examinees perceived the procedure predicts future performance regardless of face validity (Smither et al., 1993).

Whereas both face validity and perceived predictive validity are conceptually distinct, their empirical relationship is less clear. Although it is intuitively plausible to expect face validity to be highly correlated with perceived predictive validity, there has been little evidence demonstrating the correlation. The author located only one study which directly examined their empirical relationship. Smither et al. (1993) found a significant correlation of .36 ($p < .01$) between face validity and perceived predictive validity for a civil service examination. However, interpretations were problematic because the examination consisted of a variety of selection procedures (mainly paper-and-pencil measures of job knowledge and cognitive ability) and the sample of applicants were assessed for a variety of jobs ranging from entry-level to professional positions.

In the present study, the relationship between face validity and perceived predictive validity was examined separately for four different tests. Using the job of a production worker as the frame of reference, it was predicted, within each of four different tests (a situational judgement test administered either in paper-and-pencil format or in video-based format, a reading comprehension test, a personality test, and a cognitive ability test,), that:

H4: Perceived predictive validity of the test will be strongly and positively correlated with face validity of the test.

Face Validity and Method of Testing

Both work sample tests and assessment centers appear to have high face validity. Research has shown that selection procedures involving simulations elicit more favorable examinee reactions than those using paper-and-pencil measures (Dodd, 1977; Macan et al., 1944; Schmidt et al., 1977; Smither et al., 1993). Schmidt et al. (1977) reported that perceptions of work sample tests were more favorable than those of paper-and-pencil measures of cognitive ability. Dodd (1977) found that assessees have positive reactions to the face validity aspects of an assessment center. Macan et al. (1994) found that examinees perceived the assessment center as more face valid than cognitive ability tests. The high face validity and favorable examinee attitudes for work samples and assessment centers are often attributed to their realistic test situation and similarity to the target job, that is, their high simulation fidelity. Smither et al. (1993) found that procedures involving simulations were generally perceived as more favorable than paper-and-pencil measures.

However, it is not clear which aspects of these tests are responsible for the positive reactions. Previous studies comparing examinee reactions across tests (e.g., across assessment centers and cognitive ability tests) have been limited in increasing our understanding of examinee reactions due to the method-content confound across tests. By comparing two different means of measurement with test content held constant, the present study was able to examine any possible differences in reactions attributable solely to test method. The assumption that simulation fidelity or concreteness of test stimulus is positively related to examinee reactions would suggest

that in the present study, the video-based method of administering the situation judgement test would be perceived more favorably than the paper-and-pencil method even when test content remained the same because the video-based method was a more concrete representation with higher simulation fidelity than the paper-and-pencil method. There is some evidence of positive examinee reactions to a video-based method of testing (e.g., Dyer, Desmarais, Midkiff, 1993). Hence, it was predicted that

H5: Face validity of the situational judgement test will be significantly higher when administered in the video-based method than when it is administered in the paper-and-pencil method.

Within the same test, it is possible for examinees to have, simultaneously, low predictive validity perceptions and high face validity perceptions. Doing well on a test which has content related to the job tasks (i.e., a high face valid test) does not always guarantee successful job performance because successful performance has multiple determinants, many of which have little, if any, to do with test performance. Unlike face validity, perceived predictive validity is less dependent on the test content or other test characteristics. There was no clear theoretical rationale for relating differences in test methods to differences in perceived determinants of successful job performance. Hence, no formal hypothesis was formulated for any effect of method of testing on perceived predictive validity.

Subgroup Membership and Face Validity

The relationship between racial subgroup membership and reactions to selection tests is clearly an important practical issue. If examinee reactions affect subsequent examinee behaviors which are organizationally relevant, then any differential subgroup test reactions may explain some of the job performance and behavior variance or test performance associated with race. Almost all systematic differences in behaviors across racial groups have important economic and socio-political implications for the organization.

Few studies have analyzed examinee reactions by racial subgroup membership. Schmidt et al. (1977) found no Black-White differences in reactions to work sample tests and cognitive ability tests. The lack of a significant Black-White difference in attitudes toward the cognitive ability test is somewhat puzzling. Given that Blacks perform poorer than Whites on cognitive ability tests and assuming that the means of measurement (i.e., paper-and-pencil) tends to be consistent with the cultural values, beliefs, and experiences of Whites but inconsistent with those of Blacks (Helms, 1992; Goldstein, et al., 1993; Schmitt et al., 1996), one would predict that Blacks would have attitudes less favorable than Whites regarding paper-and-pencil cognitive ability tests. On the other hand, there is no reason to expect a Black-White difference in attitudes on selection procedures involving more realistic materials or concrete representations. The confound between test method and test content in Schmidt et al.'s (1977) comparison between work samples and cognitive ability test could not provide a rigorous test of the hypothesis of Black-White difference relating to method of

measurement. This hypothesis could be more directly tested in the present study because content was held constant across two different means of measurement. It was predicted that:

H6: A Race X Method interaction effect on face validity perceptions will occur. Specifically, it was predicted that for the paper-and-pencil method, the difference in face validity perceptions reported by Blacks and Whites will be greater than the difference between Black-White perceptions toward the video-based method.

The nature of the expected interaction is depicted in Figure 4.

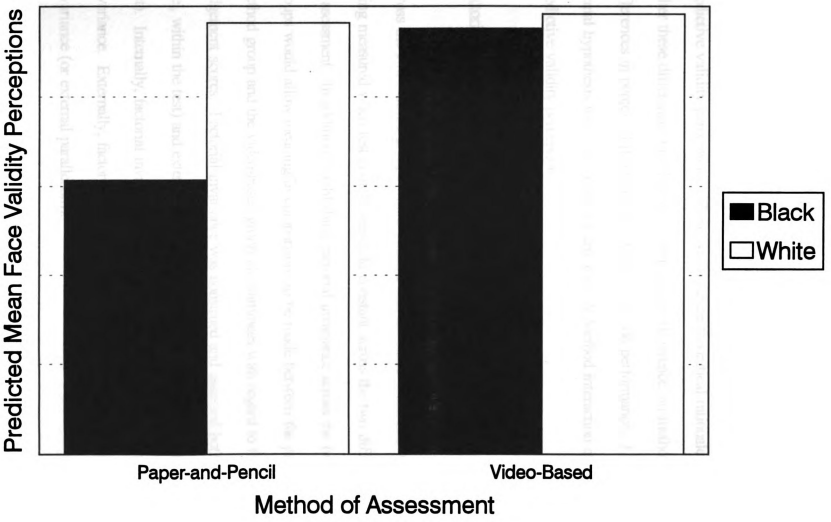

Figure 4. Hypothesis 6: Predicted Race X Method Interaction on Face Validity Perceptions of Situational Judgement Test.

Whereas racial subgroup membership was expected to interact with method of testing to affect face validity perceptions due to differential subgroup experiences with test characteristics, it was less clear if these subgroup experiences were relevant to predictive validity perceptions. There was no clear theoretical rationale for relating either these differences in subgroup experiences or differences in method of testing to differences in perceived determinants of successful job performance. Hence, no formal hypothesis was formulated for any Race X Method interaction effect on predictive validity perceptions.

Factorial Invariance across Methods of Testing

Evidence of factorial invariance of responses to the situational judgement test across the two method groups would indicate that the same constructs were indeed being measured when test content was held constant across the two different methods of asessment. In addition, establishing factorial invariance across the two method groups would allow meaningful comparisons to be made between the paper-and-pencil method group and the video-based group of examinees with regard to their situational judgement scores. Factorial invariance was construed and assessed both internally (i.e., within the test) and externally (i.e., relationships with variables external to the test). Internally, factorial invariance was construed in terms of measurement invariance. Externally, factorial invariance was construed in terms of nomological invariance (or external parallelism).

Measurement invariance exists when the numerical values across the two

groups are on the same measurement scale (Drasgow, 1984, 1987). In the absence of measurement invariance (i.e., when numerical values across groups are not on the same measurement scale), group differences in mean test scores or in patterns of correlations of the test with external variables are substantively misleading.

Nomological invariance or external parallelism across groups exists when the groups exhibit similar patterns of correlations between the test (or factors measured by the test) and external variables. To establish nomological invariance, independent established measures of personality constructs were administered in the present study for the purpose of relating them to scores on the two versions of the situational judgement test. It was anticipated that both versions would have similar patterns of correlations with the personality constructs.

METHOD

Examinees

Examinees were introductory psychology undergraduates who participated in the study for extra course credits. A series of power analyses (Cohen, 1988) was performed for each hypothesis to determine the required sample size (see Appendix A for series of power analyses). For each analysis, the power desired was .80 assuming a small effect size (see Cohen, 1988) at $\alpha = .05$. The power analyses revealed that 240 subjects were required. A total of 244 undergraduates participated in the study and 241 provided usable data (113 Blacks, 128 Whites; 63.9% females). The incomplete and unusable responses from 3 examinees were excluded from all analyses performed.

Development of Situational Judgement Test

The video-based version of the situational judgement test used in the present study was a pilot version of a video-based situations assessment test developed by a large US-based human resources consultancy firm. The test was developed by the firm as part of a comprehensive test battery for a consortium. The simulation focused on two broad functional areas namely, work habits and interpersonal skills. Each area was defined in terms of two performance factors. Work habits was defined in terms of work commitment and work quality. Interpersonal skills was defined in terms of conflict management and empathy. The videotape included one practice video vignette

and 12 actual video vignettes spanning a range of common situations likely to be encountered in today's semiskilled and skilled blue collar work place. Each vignette depicted employees interacting on the job and described an interpersonal or work-related problem for one of the employees. At the end of each vignette, examinees were asked what action the employee should take to resolve the problem. A series of possible responses (ranging 9 to 14 responses per vignette) was presented in written form on the answer booklet. For each possible response, examinees were asked to rate its appropriateness on a 6-point rating scale from very ineffective to very effective. The pilot version of the test had a total of 126 items. On the basis of an item content analyis, the human resource experts at the consultancy firm edited the test and produced a final version with a total of 63 items measuring the four a priori factors, namely, work commitment (11 items), work quality (19 items), conflict management (17 items), and empathy (16 items). The pilot version of the test was administered in the present study because the final edited version was not available at the time of study. However, only the 63 items identified in the final version of the test were used in the computation of the total situational judgement score and the analyses involving the four a priori factors.

The consultancy firm developed a rational scoring key from the ratings of 25 job content experts. Each point on the rating scale was assigned a score of 0, 1, or 2 according to the percentage of endorsement by the experts. A score of 2 was assigned when endorsement was 50% or greater, 1 when endorsement was 25% to 49.99%, and 0 when endorsement was less than 25%.

Based on the written script of the videotape which described the essential visual elements of the vignette, and the narrator's speech and dialogue between the characters both in verbatim, the present author developed a paper-and-pencil format of the test. In this paper-and-pencil measure, each of the vignettes (1 practice and 12 actual vignettes) was presented in written form. The written vignette was described in the third-person perspective (as opposed to a dialogue) similar in form to the typical paper-and-pencil type of situational judgement test used in previous research (e.g., Motowidlo et al., 1990; Pulakos et al., 1994). The substantive content of each written vignette was identical to the corresponding video vignette. After reading each vignette, examinees gave their ratings on an answer booklet similar to the one used in the video-based method containing the same response items. The scoring key for the paper-and-pencil version of the test was identical to the one used in the video-based version. The video-based administration and the paper-and-pencil administration each had a total testing time lasting 45 minutes. Appendix B presents an example of the vignettes and possible responses used in the paper-and-pencil method.

Measures of Examinee Test Reactions

Face validity and predictive validity perceptions were each assessed by a 5-item measure adapted from part of a questionnaire used in Smither et al. (1993). To provide a frame of reference, examinees were asked to give ratings on the items concerning relationships between the test and the job of a production worker working in a team-based situation. It was further stated that to do the job well, the worker had

to be both technically competent and able to relate to others effectively. Ratings were anchored on a 6-point Likert-type scale from strongly disagree to strongly agree. The questionnaire is shown in Appendix C.

Reading Comprehension, Cognitive Ability, and Personality Tests

Three widely used paper-and-pencil measures of established psychological constructs were administered to all examinees. Reading comprehension was assessed using the Comprehension subtest of the Nelson-Denny Reading Test, (Form G, Brown, Bennet, & Hanna, 1993). The test was developed for use with high school and college students and it has been widely used in psychology and education for the assessment of reading comprehension. Form G (published in 1993) is one of the two parallel forms in the fifth edition of the test that was published originally in 1929. The comprehension subtest is a multiple-choice format test in which examinees read 8 passages and respond to a total of 36 five-answer multiple choice questions. Administration time is 20 minutes. The test-retest reliabilities of the comprehension subtest reported in the test manual ranged from .75 to .82.

Cognitive ability was assessed using the Wonderlic Personnel Test (Wonderlic & Assoc., 1984). The Wonderlic test is a general cognitive test for industrial use (for reviews, see Schmidt, 1985; Schoenfeldt, 1985). It is a 12-minute test consisting of 50 items with a variety of verbal, numerical, and some spatial content, and it yields a single total score. Test-retest reliabilities ranged from .70s to .90s.

Personality constructs assessed were the "Big-five" dimensions measured using

the NEO-FFI (Costa & McCrae, 1992), a short version (i.e., 60 items) of the NEO-PI (Costa & McRae, 1985). The dimensions assessed by the test are non-clinical constructs and include conscientiousness, agreeableness, neuroticism, openness to experience, and extraversion. The test contains a total of 60 items each scored on a 5-point Likert-type scale ranging from strongly disagree to strongly agree. Each of the 5 dimensions is measured by 12 items. The time for completion of the test is 30 minutes. Evidence of criterion-related validity and construct validity for the NEO-PI have been documented in a review by Digman (1990) and reported in Costa and McCrae (1992).

Design

The study employed primarily a 2 X 2 between-subjects factorial design with performance on the situational judgement test and examinee test reactions (face validity and perceived predictive validity) as the dependent variables. The two independent variables were Race (Blacks vs. Whites) and Method (video-based vs. paper-and-pencil). Assignment of examinees to the Method condition was random with the restriction that examinees in the same testing session were administered the same method. The number of examinees per condition was approximately equal (Black-Video = 51, Black-Paper = 62, White-Video = 69, White-Paper = 59). The paper-and-pencil measures of reading comprehension, cognitive ability, and the Big-Five personality constructs were administered to all examinees.

Procedure

Examinees were tested in a classroom setting in groups ranging between 8 and 19 individuals. In the video-based method condition, the video vignettes were presented on a 25" television positioned in a manner such that all examinees could watch and listen to the videotape clearly. Instructions for the test were given on the videotape by a narrator. The instructions began with an example vignette as a practice item. The narrator first described the setting of the work situation and introduced the characters. The vignette was then presented. At the end of the vignette, the video frame froze and examinees were asked to open the answer booklet to the example situation section and indicate the effectiveness of each possible response described in written form using the 6-point rating scale. Examinees had 2 minutes to complete ratings for all the responses pertaining to the vignette. After clarifying any questions regarding the manner of completing the test, the actual test began. There were a total of 12 video vignettes on the actual test and each vignette was preceded by a narrator introduction. Examinees had 2 minutes to complete ratings for the associated responses. After the video-based test which lasted approximately 45 minutes, examinees were asked to complete a questionnaire regarding their perceptions of the test. The questionnaire was the examinee attitudes measure which consisted of the 5 items assessing face validity and the 5 items assessing perceived predictive validity. Examinees then completed a series of three paper-and-pencil measures including the Wonderlic Personnel Test, the Nelson-Denny Reading Test, and the NEO-FFI administered in counterbalanced order across test sessions. The same examinee

attitude questionnaire was administered following completion of each of the three paper-and-pencil measures.

In the paper-and-pencil method condition, examinees were presented with the paper-and-pencil version of the situational judgement test. The instructions for the test were written on the first page of the test booklet. The same example vignette preceded the 12 actual test vignettes. Examinees were given 45 minutes to complete the test. After the test, the rest of the session was identical to the video-based method session. Examinees completed the examinee attitudes questionnaire for the situational judgement test, and then followed by the three paper-and-pencil measures administered in counterbalanced order across test sessions with an examinee attitude questionnaire following completion of each measure. In both conditions, subjects were thoroughly debriefed and thanked for their participation. The total testing time per session for each condition was approximately 2 hours.

Analyses

Effect size estimates (d statistic) for subgroup differences in performance on the situational judgment test were computed by subtracting the majority test mean from the minority test mean and dividing the difference by the pooled standard deviation. Hence, negative effect sizes indicated that Blacks scored lower than Whites whereas positive effect sizes indicated the reverse.

Sex, Race, and Method were dummy coded (Females = 0, Males = 1; Whites = 0, Blacks = 1; paper-and-pencil = 0, video-based = 1) and the other study variables

were treated as continuous variables. Hierarchical regression analyses were used to test the interaction effects hypothesized in H1, H2, H3, and H6. Correlational analyses were used to test H4, and an independent-samples t-test was used to test H5.

Multiple-groups covariance structure modeling using LISREL 8 (Joreskog & Sorbom, 1993) was used to assess measurement invariance and nomological invariance of the situational judgement test across the two method groups. Measurement invariance was tested by simultaneously comparing confirmatory factor analytic models across groups. It is widely accepted that measurement invariance is established when the factor loading matrix is invariant across groups (Alwin & Jackson, 1981; Sorbom, 1974). A more stringent criterion for measurement invariance is when both factor loadings and error variances of measures are invariant across groups.

Nomological invariance was tested by comparing, across groups, the structural relationships between each situational judgement factor to the set of Big-Five personality factors. Nomological invariance is established when structural relationships are invariant across groups.

The fit of a model was assessed using the χ^2 statistic and a variety of fit indices. The χ^2 statistic is the most widely used measure of model fit in organizational research (James & James, 1989; Kelloway, 1996). The main disadvantage of the χ^2 is its high sensitivity to sample size such that with large sample sizes, most models will produce statistically significant χ^2 values resulting in rejection of these models even if they are theoretically reasonable. Hence, most researchers

also rely on a variety of alternate fit indices to reduce the dependence on sample size when assessing model fit. Because the various indices differ on their specific assumptions, the use of multiple indices when evaluating a model can provide convergent evidence in the assessment of model fit. In the present study, the indices used included Joreskog and Sorbom's (1989) goodness-of-fit index (GFI) and adjusted goodness-of-fit index (AGFI), Bentler's (1990) comparative fit index (CFI), Bentler and Bonett's (1980) non-normed fit index (NNFI), Joreskog and Sorbom's (1986) standardized root mean square residual (standardized RMSR), and Steiger's (1990) root mean square error of approximation (RMSEA).

Both GFI and AGFI are widely used indices of fit based on the comparison of observed and estimated covariances (see Kelloway, 1996). The AGFI is a parsimonous fit which adjusts the GFI for the degrees of freedom in the model, that is, it takes into consideration the fact that a model always increases in fit as the number of free parameters to be estimated approaches the number of independent pieces of information available for estimation. The CFI and NNFI measure how well the model fits relative to a baseline model, usually the independence (i.e., null) model. The values of GFI, AGFI, CFI, and NNFI range from 0 to 1.0 with values approaching 1.0 indicating a good fit to the data. The present study used the convention of larger than .90 as an indication of good fit.

The standardized RMSR is a measure of the average standardized residuals of the predicted covariance matrix from the observed covariance matrix. Values approaching 0 indicate a good fit to the data. The conventional value of less than .10

was used as an indication of good fit in the present study. The RMSEA is a measure of the average size of the fitted residuals per degree of freedom. Following Browne and Cudeck (1993), the present study considered a value of .05 or less as indicating a close fit; between .05 and .10 as a moderate fit; and more than .10 as a poor fit.

The χ^2 difference test ($\Delta\chi^2$), obtained by calculating the difference in the models' respective χ^2 with degrees of freedom equal to the difference in the models' respective degrees of freedom, was used to compare the statistical significance of difference in fit between nested models.

RESULTS

Table 1 presents the means, standard deviations, reliability estimates, and intercorrelations of all the study variables. The same statistics broken down by racial subgroups are reported in Appendix D. As shown in Table 1, the internal consistency reliability estimates (Cronbach's α) for the measures used in the present study were in acceptable ranges. The reliability estimates reported for the two versions of the situational judgement test are underestimates because of the multidimensional nature of the test.

An inspection of Table 1 showed bivariate support for the major hypotheses. Race was more highly correlated with situational test performance when the test was administered in the paper-and-pencil method than when administered in the video-based method. Consistent with previous research, race was correlated with reading comprehension. With regard to test reactions, face validity perceptions and predictive validity perceptions were positively correlated for each of the four different tests (i.e., situational judgement, reading comprehension, cognitive ability, personality) used in the study. For the situational judgement test, face validity perceptions were correlated with the method of assessment. Also, race was more highly correlated with face validity perceptions when the situational judgement test was administered in the paper-and-pencil method than when administered in the video-based method. Each hypothesis will be addressed directly and in a multivariate sense in the following sections.

Table 1

Means, Standard Deviations, Reliabilities, and Intercorrelations of Study Variables

	Means	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
Variables[a]																								
1. METHOD	.50	.50																						
2. SEX	.36	.48	06																					
3. RACE	.47	.50	09	09																				
4. PSJ	39.37	8.39	----	-12	-47	(64)																		
5. VSJ	42.86	7.72	----	-07	-11	----	(55)																	
6. READING	29.15	5.73	-13	04	-35	45	05	(84)																
7. CONSC	32.21	6.01	01	-05	-14	06	02	13	(80)															
8. AGREE	30.38	6.51	-08	-22	-07	34	13	09	24	(78)														

Table 1 continues

	Means	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
9. NEURO	22.14	7.99	04	-20	02	-18	10	-20	-30	-20	(84)													
10. OPEN	30.08	5.85	-03	-04	-21	24	05	14	-01	09	05	(71)												
11. EXTRAV	32.16	6.21	-07	-23	-14	11	-16	02	23	30	-16	14	(78)											
12. COGN	23.19	5.49	-06	01	-29	30	29	45	-03	-03	-04	15	-03	(88)										
13. FACE- PSJ	17.84	4.36	----	-26	-40	38	----	33	16	06	02	21	10	34	(90)									
14. PRED-PSJ	13.14	3.20	----	-02	02	03	----	12	12	-02	-07	04	-03	05	28	(75)								
15. FACE-VSJ	19.69	2.96	----	-05	-05	----	15	13	12	16	07	-08	-05	08	----	----	(78)							
16. PRED-VSJ	14.08	3.55	----	-10	00	----	14	09	08	21	06	-11	-05	04	----	----	24	(81)						
17. FACE-READ	13.46	3.58	06	02	-24	08	02	05	15	-05	-02	03	10	08	04	07	11	09	(88)					
18. PRED-READ	11.80	3.61	03	06	-12	05	13	07	09	-03	-07	01	-06	16	04	33	10	26	60	(90)				
19. FACE- NEO	16.61	3.20	-07	-07	00	04	-06	02	12	07	-13	-03	10	-07	-01	15	17	13	26	20	(76)			
20. PRED-NEO	13.23	3.92	03	-02	05	-04	24	05	08	11	-09	-01	04	06	05	42	00	35	13	45	48	(86)		

Table 1 continues

	Means	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
21. FACE-COGN	13.05	3.37	05	03	-23	03	07	09	11	-08	-07	09	-06	07	03	17	06	19	47	43	09	13	(81)	
22. PRED-COGN	11.56	3.49	-02	05	-18	03	15	08	05	-06	-01	05	-11	14	-01	41	06	40	34	55	11	37	70	(86)

[a] Variables abbreviated: METHOD=Method of assessment in situational judgement test; SEX=Sex of examinee; RACE=Race of examinee; PSJ=Paper-and-Pencil Situational Judgement; VSJ=Video-based Situational Judgement; READ=Nelson-Denny Reading Comprehension; NEO= NEO-FFI Personality; CONSC=Conscientiousness; AGREE=Agreeableness; NEURO=Neuroticism; OPEN=Openness; EXTRAV=Extraversion; FACE=Face Validity; PRED=Perceived Predictive Validity.

Note. METHOD, SEX, and RACE are dummy-coded (Video-based=0, Paper-and-Pencil=1; Female=0, Male=1; White=0, Black=1). Decimals for all correlations and reliabilities are omitted. Reliabilities are in parentheses. All reliability estimates are Cronbach's α except for COGN which is a test-retest reliability (from Wonderlic manual). Cronbach's α for PSJ and VSJ are underestimates because of the multidimensional nature of the tests. Means, standard deviations, and intercorrelations for all variables are based on $N = 241$ except for PSJ ($N = 121$), FACE-PSJ ($N = 121$), PRED-PSJ ($N = 121$), VSJ ($N = 120$), FACE-VSJ ($N = 120$), and PRED-VSJ ($N = 120$).

Relationships between Race, Reading Comprehension, Method of Assessment, and Performance on Situational Judgement Test (Hypotheses 1, 2, and 3)

Table 2 presents the hierarchical regression analyses performed to test H1, H2, and H3. H1 predicted a Race X Method interaction on performance on the situational judgement test such that the Black-White difference using the video-based method of testing will be smaller than that using the paper-and-pencil method. As shown in Table 2, Race and Method were entered as a single block in step 1 of the regression of test performance on race and method of assessment. These effects accounted for 12% of the test variance, $p < .05$. The Race X Method product term, which represented the Race X Method interaction, was entered in step 2 of the regression. Entering the Race X Method interaction term resulted in a significant increase in variance accounted for, $\Delta R^2 = .04$, $\Delta df = 1$, $p < .05$.

Table 2

Summary of Hierarchical Regressions of Situational Judgement Test Performance on Race, Reading Comprehension, and Method of Assessment (N = 241)

Hypothesis and Predictors	R^2	df	ΔR^2	Δdf	ΔF
Hypothesis 1					
Step 1. Race	.12	2			17.40*
Method					
Step 2. Race X Method	.16	3	.04	1	9.78*
Hypothesis 2					
Step 1. Reading	.12	2			16.50*
Method					
Step 2. Reading X Method	.15	3	.03	1	7.48*
Hypothesis 3					
Step 1. Race	.19	4			13.85*
Reading					
Method					
Reading X Method					
Step 2. Race X Method	.20	5	.01	1	4.21*

* $p < .05$.

Figure 5 depicts the nature of the interaction in terms of differences in subgroup mean performance. As shown in the figure, the Black-White difference in test performance was greater in the paper-and-pencil method than in the video-based method. To assess the practical significance of the statistically significant Race X Method interaction, effect sizes for subgroup observed mean differences were computed using the d statistic. A substantial Black-White difference in performance of almost one standard deviation favoring Whites was found on the paper-and-pencil version of the situational judgement test, d = -.95. The Black-White difference was reduced substantially to about one-fifth of a standard deviation in the video-based version of the test, d = -.21. Hence, H1 was supported.

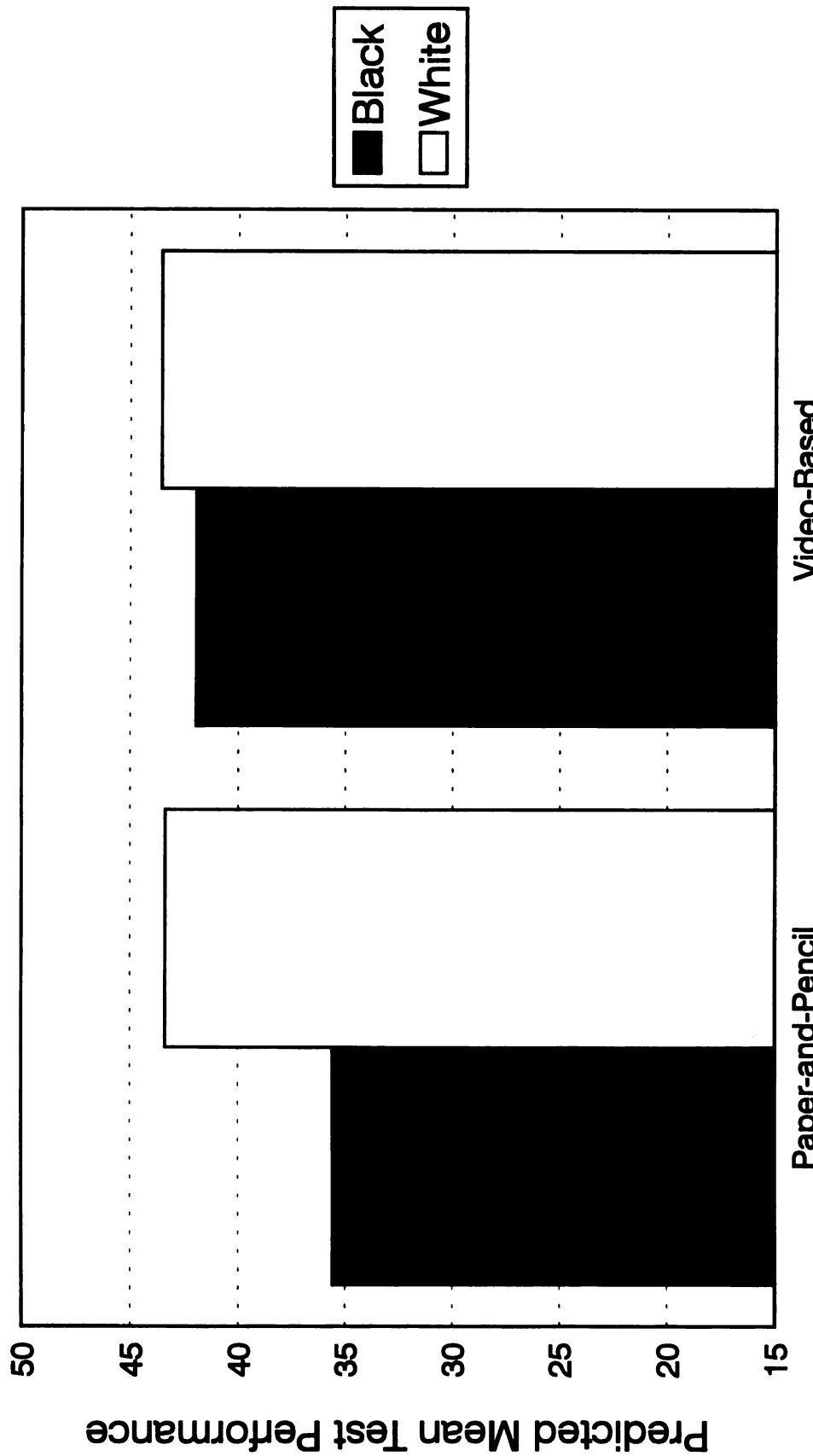

Figure 5. Race X Method Interaction on Situational Judgement Test Performance.

H2 predicted that a Method X Reading Comprehension interaction effect on situational judgement test performance such that performance will be positively and significantly correlated with reading comprehension ability in the paper-and-pencil method of testing whereas no significant correlation between test performance and reading comprehension ability will occur in the video-based method. As shown in Table 2, entering Method and Reading Comprehension as a single block in step 1 of the regression of test performance on these factors accounted for 12% of the variance, $p < .05$. The Method X Reading Comprehension interaction term was entered in step 2 which resulted in a significant increase in variance accounted for, $\Delta R^2 = .03$, $\Delta df = 1$, $p < .05$. A plot of the interaction (Cohen & Cohen 1988) as depicted in Figure 6 showed that test performance and reading comprehension were positively correlated in the paper-and-pencil method of testing but they were nearly uncorrelated in the video-based method. Hence, H2 was supported.

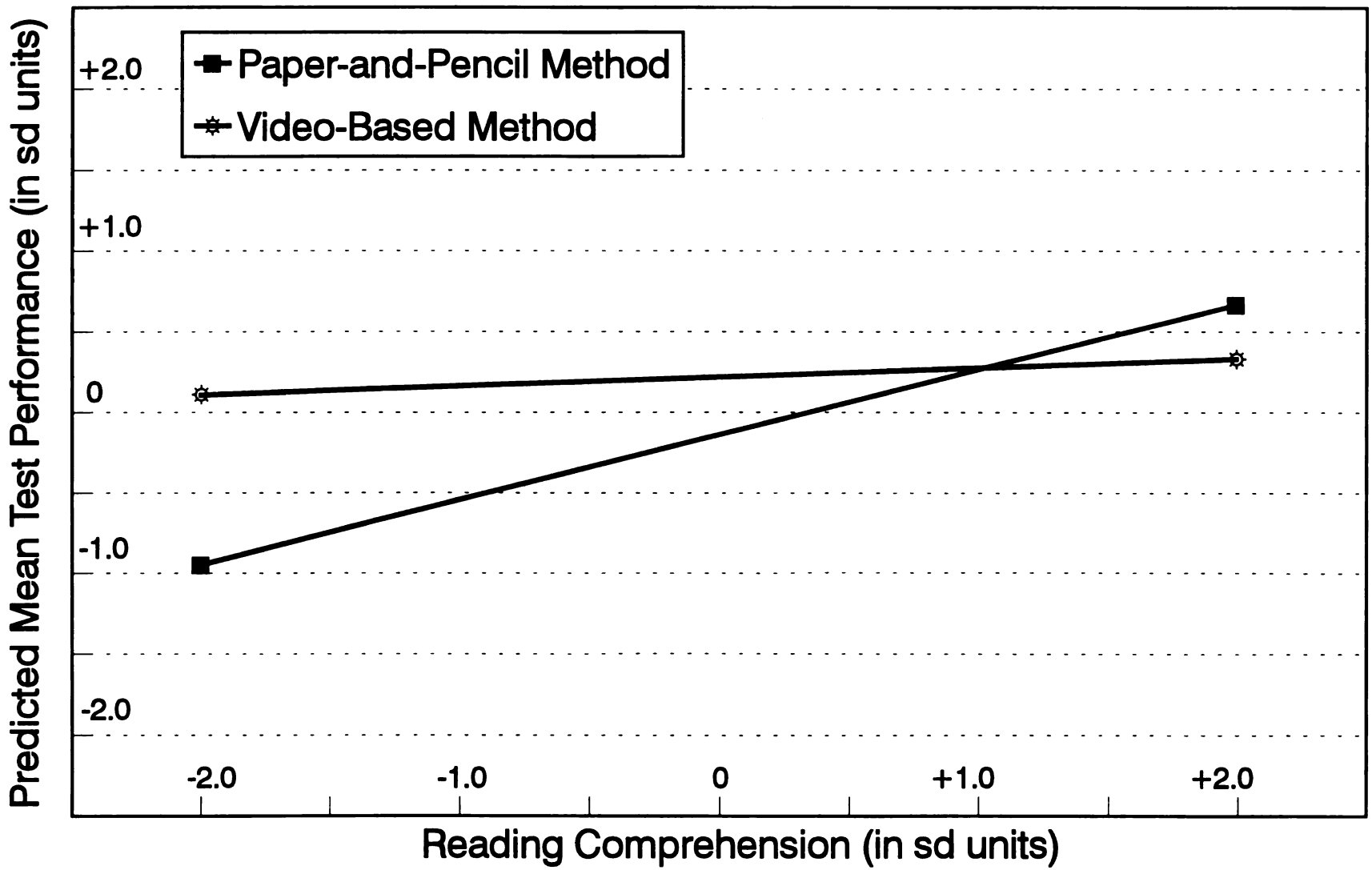

Figure 6. Method X Reading Comprehension Interaction on Situational Judgement Test Performance.

H3 predicted that the Race X Method interaction effect on situational judgement test performance would diminish after controlling for the effect of the Method X Reading Comprehension interaction. As shown in Table 2, Race, Reading Comprehension, Method, and Method X Reading Comprehension interaction were entered as a single block in step 1 of the regression of test performance on race, reading comprehension, and method of assessment. The block accounted for 19% of the variance, $p < .05$. Entering the Race X Method interaction term in step 2 provided a significant but small increase in variance accounted for, $\Delta R^2 = .01$, $\Delta df = 1$, $p < .05$. The proportion of variance in test performance accounted for by the Race X Method interaction obtained in H1 diminished substantially from 4% to a small (though still statistically significant, $p < .05$) 1% once the effect of Method X Reading Comprehension on test performance was controlled. Hence, H3 was supported.

Figure 7 depicts the nature of the Race X Method interaction after controlling for the effect of Method X Reading Comprehension interaction on test performance. Compared to Figure 5, Figure 7 shows that the Race X Method interaction effect was dampened to some extent after controlling for the effect of Method X Reading Comprehension interaction.

In summary, the regression analyses provided support for the first three hypotheses. There was a Race X Method interaction effect on situational judgement test performance such that the Black-White performance difference (favoring Whites) was substantially smaller in the video-based method of testing than in the paper-and-pencil method. A Method X Reading Comprehension interaction also existed such

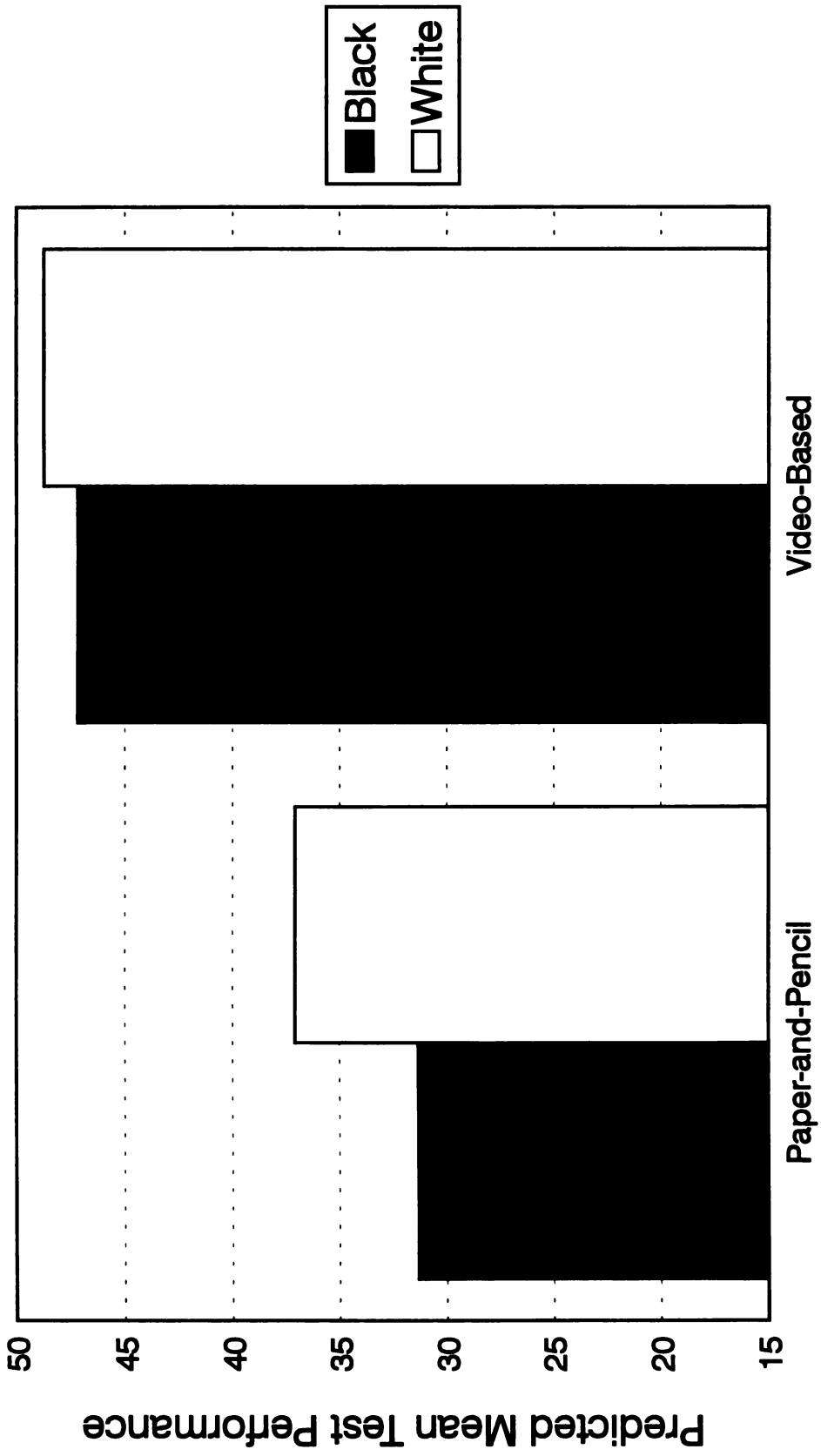

Figure 7. Race X Method Interaction on Situational Judgement Test Performance after Controlling for Effect of Method X Reading Comprehension Interaction.

that test performance was positively correlated with reading comprehension ability in the paper-and-pencil method but that they were nearly uncorrelated in the video-based method. As shown in the regression results for H3, this Method X Reading Comprehension interaction accounted for a substantial portion of the Race X Method interaction effect on test performance.

Factorial Invariance across Method Groups

Table 3 presents the means, standard deviations, and both observed and corrected (for scale unreliability) intercorrelations of the four a priori scales on the situational judgement test, broken down by method groups. Not surprisingly, internal consistency estimates of reliabilities (Cronbach's α) were low due to the relatively small number of items on each scale and the dichotomous (with a few trichotomous) scoring of the items. However, scale reliabilities were substantially higher than inter-scale correlations which provided some preliminary evidence for discriminant validity. Inter-scale correlations remained low, relative to scale reliabilities, even after correcting for unreliability in each scale. Multiple-group covariance structure analysis was used to provide a more rigorous test for the discriminant validity of the four a priori scales and to assess factorial invariance across method groups.

Table 3

Means, Standard Deviations, Reliabilities, and Intercorrelations (observed and corrected) of Situational Judgement Scales broken down by Method Groups

	Means	SD	1	2	3	4
Situational Judgement Scales						
Paper-and-Pencil (N = 121)						
1. Conflict	11.84	3.81	(.46)	.49	.23	.36
2. Empathy	9.47	3.65	.24	(.53)	.70	.28
3. Quality	10.09	2.50	.08	.26	(.26)	.27
4. Commitment	7.98	3.54	.18	.15	.10	(.53)
Video-Based (N = 120)						
1. Conflict	12.29	3.64	(.40)	.48	.49	.30
2. Empathy	10.44	3.03	.15	(.24)	.85	.49
3. Quality	10.74	2.78	.13	.27	(.42)	.22
4. Commitment	9.38	3.33	.12	.15	.09	(.39)

Note. Cronbach's α reliabilities are in parentheses. Observed correlations are below diagonals and corrected (for unreliability) correlations are above diagonals.

Measurement Invariance. As described earlier, factorial invariance referred to both measurement invariance and nomological invariance. To formulate measurement models for the test of measurement invariance, items within each of the four scales were first randomly sorted into three sets comprised of approximately equal numbers of items. Item scores were unit-weighted and summed within each set to create three trait indicators (also known as observed indicators) for each latent trait variable purportedly measured by each scale (i.e., each situational judgement factor), giving a total of 12 trait indicators. A factor loading was arbitrarily set to 1.0 for one of the three indicators for each latent trait variable in order to scale that latent trait variable (Bollen, 1989). Appendix E presents the 12 X 12 observed covariance matrix among trait indicators for each of the two method groups.

Table 4 presents the fit indices associated with the series of nested confirmatory factor analytic models fit to the two observed covariance matrices. Also presented in this table are chi-square difference tests associated with relevant model comparisons.

Table 4

Fit Indices Associated with Multiple-Group Confirmatory Factor Analytic Models Tested in Assessment of Measurement Invariance of Situational Judgement Scores across Paper-and-Pencil Method Group (N = 121) and Video-Based Method Group (N = 120)

Model Specifications Across Groups	χ^2	df	Model Comparison	$\Delta\chi^2$	Δdf	GFI	AGFI	CFI	NNFI	std RMSR	RMSEA
M1. Single General Factor. Free factor loadings and error variances.	181.40*	109				.88	.91	.66	.59	.09	.05
M2. Four Correlated Factors. Free factor covariances, factor loadings, and error variances.	123.48	100	M1 vs M2	57.92*	9	.91	.93	.89	.86	.08	.03
M3. Four Correlated Factors. Equal factor loadings; free factor covariances and error variances	133.68	108	M2 vs M3	10.20	8	.90	.93	.88	.85	.09	.03

Table 4 continues

Model Specifications Across Groups	χ^2	df	Model Comparison	$\Delta\chi^2$	Δdf	GFI	AGFI	CFI	NNFI	std RMSR	RMSEA
M4. Four Correlated Factors. Equal factor loadings and error variances; free factor covariances.	140.71	120	M3 vs M4	7.03	12	.90	.94	.90	.89	.09	.03
M5. Four Correlated Factors. Equal factor loadings, error variances, and factor covariances.	142.72	126	M4 vs M5	2.01	6	.90	.94	.92	.92	.09	.02
			M2 vs M5	19.24	26						

* $p < .05$.

A single general factor model in which factor loadings and error variances were freely estimated across method groups (Model M1) was first fit to the covariance matrices as the baseline measurement model. The single general factor model provided a marginal fit to the data, $\chi^2 = 181.40$, $df = 109$, $p < .05$, GFI = .88, AGFI = .91, CFI = .66, NNFI = .59, standardized RMSR = .09, RMSEA = .05. Hence, there was no strong evidence of unidimensionalty in the situational judgement test.

A four factor model in which factor covariance, factor loadings, and error variances were freely estimated across method groups (Model M2) was next fit to the data. The model provided a significant increase in fit over the single factor model, $\Delta\chi^2 = 57.92$, $\Delta df = 9$, $p < .05$, and a reasonable fit to the data as indicated by the fit indices. To test for measurement invariance across method groups, Model M2 was compared to a more parsimonous model (i.e., with higher degrees of freedom) in which factor loadings were constrained to be equal across groups. The more parsimonous model (i.e., Model M3) continued to provide a reasonable fit to the data and the decrease in model fit from Model M2 to Model M3 was nonsignificant, $\Delta\chi^2 = 10.20$, $\Delta df = 8$, n.s. Hence, equality of factor loadings across method groups was established.

Model M3 was compared to Model M4, a yet more parsimonous model in which both factor loadings and error variances were constrained to be equal across groups. Model M4 provided a good fit as indicated by the fit indices and the decrease in model fit, as measured by the χ^2 difference test, from Model M3 to Model M4 was nonsignificant, $\Delta\chi^2 = 7.03$, $\Delta df = 12$, n.s. That is, using the stringent criterion of both

equal factor loadings and equal error variances, measurement invariance across the two method groups was established.

To examine the structural aspects of the confirmatory factor analytic models (i.e., the relationships among latent trait variables), Model M4 was compared to Model M5 in which between-group equality constraints were imposed on factor covariances. In Model M5, the six factor covariances were constrained to be equal across the two method groups. That is, Model M4 and Model M5 differed only with respect to structural relations among the latent trait variables; for each model, the factor loadings and error variances were constrained to be equal across the two method groups.

Model M5 provided a good fit to the data, $\chi^2 = 142.72$, $df = 126$, n.s., GFI = .90, AGFI = .94, CFI = .92, NNFI = .92, standardized RMSR = .09, RMSEA = .02. The decrease in model fit from Model M4 to Model M5 was nonsignificant, $\Delta\chi^2 = 2.01$, $\Delta df = 6$, n.s. Comparison between Model M5 and Model M2 also revealed that as a whole, none of the equality constraints on factor covariances, factor loadings, and error variances significantly decreased model fit. Hence, Model M5 was selected as the most adequate measurement model. Figure 8 depicts Model M5 with its associated common metric factor loadings and factor correlations. All factor loadings were statistically significant, $p < .05$. Of the six factor correlations, five were statistically significant, $p < .05$. Full measurement invariance across method groups (i.e., full internal factorial invariance) was established in terms of error variances, factor loadings and factor covariances.

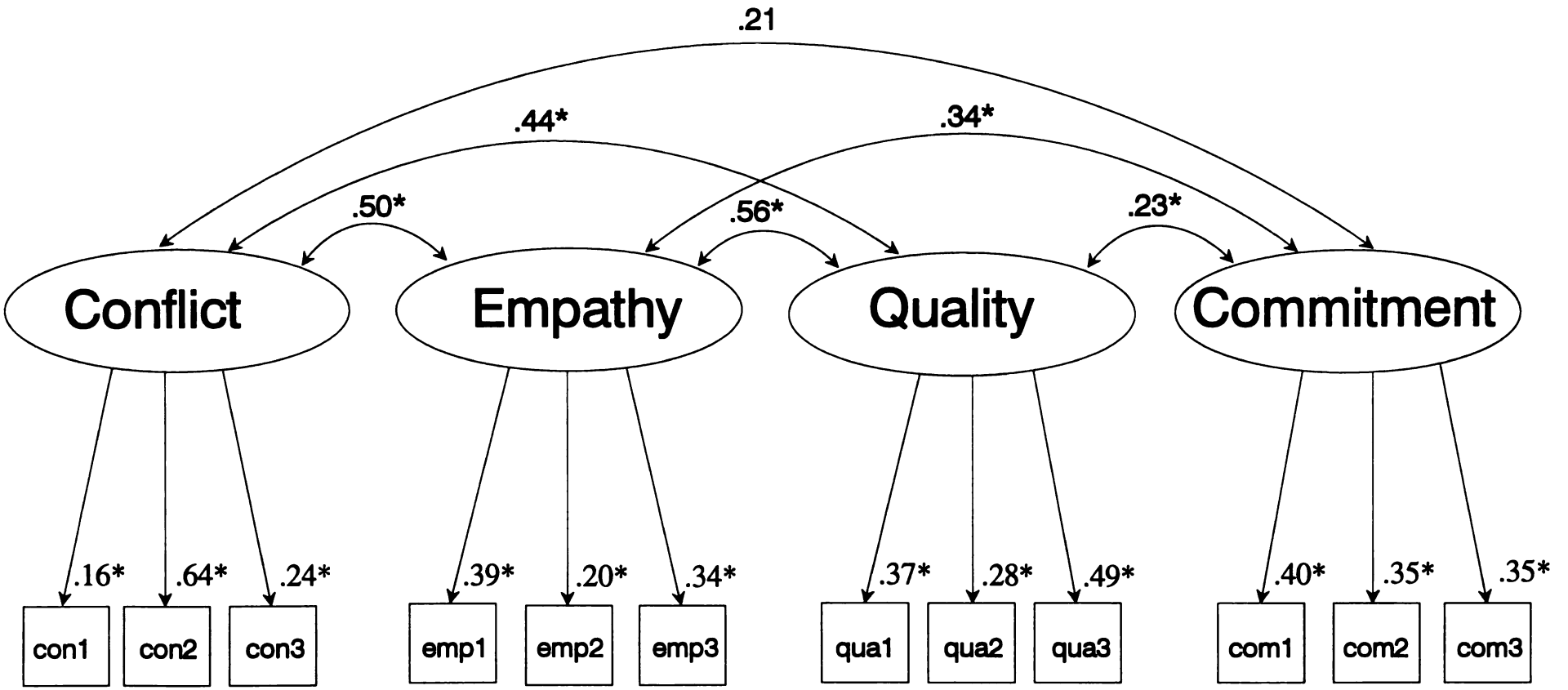

Figure 8. Confirmatory Factor Analytic Model with Associated Common Metric Standardized Factor Loadings and Factor Correlations for Both Method Groups (* $p < .05$).

Nomological Invariance. Table 5 presents, for each method group, the means, standard deviations, and intercorrelations between the 12 situational judgement indicators and the 5 personality indicators.

Table 5

Means, Standard Deviations, and Intercorrelations between Situational Judgement Indicators and Personality Indicators broken down by Method Groups

			Situational Judgement Indicators											
	Means	SD	Con1	Con2	Con3	Emp1	Emp2	Emp3	Qua1	Qua2	Qua3	Com1	Com2	Com3
Personality Indicators														
Paper-and-Pencil (N=121)														
CONSC	32.27	6.36	.18	-.16	.14	-.08	-.16	-.09	.09	.00	.06	.02	.07	.14
AGREE	29.87	6.70	.26	.20	.26	.20	.14	.03	.05	.07	.10	.21	.06	.11
NEURO	22.49	8.18	-.01	-.02	-.05	-.08	-.09	-.02	-.16	-.11	-.15	-.12	-.07	-.09
OPEN	29.92	5.94	.16	.16	.22	.00	.08	.16	.13	-.07	.09	.19	.08	-.02
EXTRAV	31.74	5.98	.13	.10	.19	.02	.01	-.01	.05	.05	.05	.04	-.07	-.03
Means			3.85	2.83	5.16	4.40	2.70	2.37	4.27	3.05	2.77	3.07	2.33	2.57

Table 5 continues

			Situational Judgement Indicators											
	Means	SD	Con1	Con2	Con3	Emp1	Emp2	Emp3	Qua1	Qua2	Qua3	Com1	Com2	Com3
SD			1.99	1.32	2.00	1.69	1.54	1.57	1.44	1.13	1.21	1.74	1.66	1.51
Video-Based (N=120)														
CONSC	32.14	5.67	.07	-.07	-.02	.06	-.08	-.10	-.03	.03	.10	.06	.10	-.04
AGREE	30.89	6.30	.10	-.08	.13	.10	-.04	-.03	.05	.09	.09	.15	.05	.00
NEURO	21.78	7.82	-.09	.14	.02	-.01	-.06	.10	.09	.13	-.05	.02	.09	.13
OPEN	30.24	5.78	.01	.11	.25	.08	.00	-.11	-.06	-.18	-.06	.21	-.05	-.01
EXTRAV	32.58	6.43	-.03	-.22	-.11	-.07	-.05	-.03	-.20	-.13	-.16	.10	.10	-.03
Means			4.14	3.10	5.05	4.80	2.76	2.88	4.33	3.42	2.98	3.23	3.12	3.03
SD			2.06	1.31	1.85	1.62	1.40	1.65	1.62	1.09	1.17	1.47	1.64	1.61

Note. Variables abbreviated: Con=Conflict; Emp=Empathy; Qua=Quality; Com=Committment; CONSC=Conscientiousness; AGREE=Agreeableness; NEURO=Neuroticism; OPEN=Openness; EXTRAV=Extraversion.

The present author had planned to use a multiple-group covariance structure analysis approach to testing equality of structural relationships between latent variables across groups (Joreskog & Sorbom, 1993) to assess nomological invariance (external parallelism) of the four situational judgement factors (in reference to the Big-Five personality factors) across method groups. Full nomological invariance is achieved when both equality of parameter estimates and equality of errors in each of the structural equations relating the respective situational judgement factor to the set of personality factors are established across method groups.

However, as shown in Table 5, the observed correlations between the situational judgement indicators and the personality indicators were trivial, fluctuating around 0. Contrary to the author's expectation, it appeared that the personality factors measured in the present study were not related to the situational judgement factors. Therefore, it was not meaningful to test for external parallelism using the five personality factors as external reference variables. A multiple-group covariance structure analysis was attempted, but failed to reject a model specifying between-group equality in structural parameters. The failure to reject was due to the lack of correlation between situational judgement factors and the external reference variables used for both method groups, and not because of a between-group similarity in the patterns of external correlation (i.e., external parallelism). Because of the low correlations between situational judgement scales and personality measures, it was found that structural parameter estimates in both method groups were trivial.

Three nested models were fit to the 17 X 17 observed covariance matrix

relating the 12 indicators for the situational judgement factors and 5 indicators for personality factors for each method group (covariance matrices are reported in Appendix F). For all three models, measurement aspects were held constant so that effects of structural and measurement differences were not confounded in model comparisons. Constraining measurement aspects also resulted in the comparison of more parsimonous models by reducing the number of parameters to be estimated simultaneously.

For measurement aspects of the four situational judgement factors, factor covariances, factor loadings and error variances of observed indicators were constrained to be equal across method groups (i.e., the measurement model specified by Model M5). For measurement aspects of the five personality factors, the error variances of observed indicators are not identified parameters and they cannot be estimated because there was only one observed indicator (i.e., Big-Five sub-scale) per factor. Rather than assuming that the indicators were infallible measures by fixing error variances to zero, the identification problem was solved by fixing the error variance of each indicator to a value derived from its internal consistency estimate of reliability r_{xx} (Cronbach's α), using the formula

$$\sigma_e^2 = (1-r_{xx})\sigma_x^2 \qquad (1)$$

where σ_e^2 = error variance of indicator and σ_x^2 = variance of indicator (Joreskog and Sorbom, 1993).

Model N1 freely estimated across method groups both the structural parameters and error terms in each of the four structural equations relating the respective

situational judgement factor to the set of five personality factors. The model provided a good fit to the data, $\chi^2 = 164.84$, $df = 217$, n.s., GFI = .92, AGFI = .94. The model was compared to the more parsimonous Model N2 which similarly allowed error terms to freely vary but specified structural parameter estimates to be invariant across method groups. A χ^2 difference test showed that decrease in model fit from Model N1 to Model N2 was nonsignificant, $\Delta\chi^2 = .72$, $\Delta df = 20$, n.s. Hence, structural parameter estimates did not differ significantly across method groups. Model N2 was compared to Model N3 which specified both structural parameter estimates and error terms structural equations to be invariant across method groups. The decrease in model fit from Model N2 to Model N3 was nonsignificant, $\Delta\chi^2 = 2.08$, $\Delta df = 4$, n.s. That is, error terms in structural equations did not differ significantly across groups.

However, for all three models, an inspection of the common metric standardized regressions of each situational judgement factor on the five personality factors revealed trivial parameter estimates fluctuating around 0, within the range between -.07 and +.04. Therefore, whereas the nested model comparisons indicated equality of structural equations relating the respective situational judgement factor to the set of personality factors, the equality should not be construed as evidence for nomological invariance across method groups (i.e., external factorial invariance/parallelism). Instead, the equality of structural regressions was a result of near-zero correlations between situational judgement factors and the external reference variables (i.e., the Big-Five personality measures) selected for the assessment of external parallelism.

Effects of Method of Assessment on Differential Subgroup Performance on Individual Situational Judgement Constructs

The establishment of factorial invariance of responses to the situational judgement test in terms of full measurement invariance across method groups supported the meaningfulness of between-method comparisons of subgroup performance at the level of individual constructs measured by the test. Factor scores for each of the four situational judgement factors were computed for all examinees based on the factor loadings in the measurement model (Model M5). Because the factors are latent variables free of measurement errors in the observed indicators, comparisons of Black-White differences in factor scores provide more accurate estimates (i.e., disattenuated for unreliability in measures) of the effect of method of assessment on adverse impact in the situational judgement test.

Table 6 presents, for each of the four situational judgement factors, the subgroup factor means, standard deviations, and associated d statistics for each of the two methods of assessment. As shown in the table, the paper-and-pencil method produced substantial Black-White differences in performance favoring Whites on each of the four constructs as indicated by the d statistic (Conflict = -.70; Empathy = -.43; Quality = -.35; Commitment = -.63). These Black-White differences were substantially reduced in the video-based method (Conflict = .02; Empathy = -.18; Quality = .06; Commitment = -.36), with d differences across methods ranging from .27 to .72. In fact, in the video-based method, Black-White differences were not statistically significant for any of the four factors.

Table 6

Situational Judgement Factors: Subgroup Means, Standard Deviations, and Associated Effect Sizes for Paper-and-Pencil Method and Video-Based Method of Assessment

	Paper-and-Pencil Method				Video-Based Method				
	N	Means	SD	d statistic	N	Means	SD	d statistic	Difference in d
Situational Judgement Factor									
Conflict									
Black	62	4.43	1.66		51	5.30	1.54		
White	59	5.62	1.43		69	5.27	1.69		
Total	121	5.01	1.70	-.70*	120	5.29	1.62	.02	.72
Empathy									
Black	62	4.43	1.83		51	5.14	1.44		

Table 6 continues

	Paper-and-Pencil Method				Video-Based Method				
	N	Means	SD	d statistic	N	Means	SD	d statistic	Difference in d
White	59	5.21	1.76		69	5.42	1.58		
Total	121	4.81	1.83	-.43*	120	5.30	1.52	-.18	.25
Quality									
Black	62	4.42	1.15		51	4.92	1.28		
White	59	4.83	1.19		69	4.84	1.38		
Total	121	4.62	1.18	-.35	120	4.88	1.33	.06	.41
Commitment									
Black	62	3.58	1.87		51	4.49	1.53		
White	59	4.74	1.63		69	5.12	1.87		
Total	121	4.15	1.84	-.63*	120	4.83	1.76	-.36	.27

* $p < .05$.

To summarize, nomological invariance of responses to the situational judgement test across the two method groups could not be tested because of near-zero correlations between situational judgement factors and the external reference personality variables. However, factorial invariance in terms of full measurement invariance across methods was established. Measurement invariance supported the meaningfulness of between-method comparisons of racial subgroup performance at the level of individual constructs disattenuated for measurement errors. For each construct, there was a large Black-White performance difference favoring Whites in the paper-and-pencil method. These performance differences were substantially reduced in the video-based method.

Face Validity and Predictive Validity Perceptions

H4 predicted that for each of the four different tests administered in the present study predictive validity perceptions will be strongly and positively correlated with face validity perceptions. Results showed that correlations between the two perceptions were significant ($p < .05$), positive, and substantial for all four tests (paper-and-pencil situational judgement, $r = .28$, $N = 121$; video-based situational judgement, $r = .24$, $N = 120$; reading comprehension, $r = .60$, $N = 241$, personality, $r = .48$, $N = 241$; cognitive ability, $r = .70$, $N = 241$). For each test, the correlation between the two types of perceptions was substantially lower than the reliability estimates (Cronbach's α) of the respective perception measures (paper-and-pencil situational judgement, Face $r_{xx} = .90$, Predictive $r_{xx} = .75$; video-based situational

judgement, Face r_{xx} = .78, Predictive r_{xx} = .81; reading comprehension, Face r_{xx} = .88, Predictive r_{xx} = .90; personality, Face r_{xx} = .76, Predictive r_{xx} = .86; cognitive ability, Face r_{xx} = .81, Predictive r_{xx} = .86). This provided evidence of discriminant validity for the two types of perceptions. H4 was supported.

Face Validity and Method of Testing

H5 predicted that face validity perceptions of the situational judgement test will be significantly higher when administered in the video-based method than when it is administered in the paper-and-pencil method. Results of an independent sample t-test supported the hypothesis; the video-based method received significantly higher mean face validity ratings (M = 19.69, SD = 2.96, N = 120) than the paper-and-pencil method (M = 17.84, SD = 4.36, N = 121), t (237) = 3.87, $p < .05$.

Subgroup Membership and Face Validity

H6 predicted that the difference in face validity perceptions on the situational judgement test reported by Blacks and Whites will be greater in the paper-and-pencil method than in the video-based method. To test this Race X Method interaction, a hierarchical regression of face validity perceptions was performed. As shown in Table 7, Race and Method were entered as a single block in step 1 of the regression and accounted for 12% of the variance in perceptions, $p < .05$. Entering the Race X Method interaction term in step 2 of the regression resulted in a significant increase in variance accounted for, $\Delta R^2 = .04$, $\Delta df = 1$, $p < .05$.

Table 7

Hierarchical Regressions for Face Validity Perceptions and Situational Judgement Test Performance (N = 241)

Criteria and Predictors		R^2	df	ΔR^2	Δdf	ΔF
Face Validity (Hypothesis 6)						
Step 1.	Race	.120	2			16.52*
	Method					
Step 2.	Race X Method	.160	3	.040	1	12.13*
Test Performance						
Step 1.	Race	.219	5			13.20*
	Reading					
	Method					
	Reading X Method					
	Face Validity					
Step 2.	Race X Method	.227	6	.008	1	2.50

* $p < .05$.

Figure 9 depicts the nature of the interaction in terms of differences in subgroup mean perceptions. As shown in the figure, the Black-White difference in face validity perceptions on the situational judgement test was greater in the paper-and-pencil method than in the video-based method. To assess the practical significance of the statistically significant Race X Method interaction, effect sizes for subgroup differences were computed using the d statistic. A substantial Black-White difference in perceptions of four-fifths of a standard deviation with Whites reporting higher face validity was found on the paper-and-pencil version of the situational judgement test, d = -.80. The Black-White difference in perceptions was reduced substantially to a practically trivial one-nineth of a standard deviation in the video-based version of the test, d = -.11. Hence, H6 was supported.

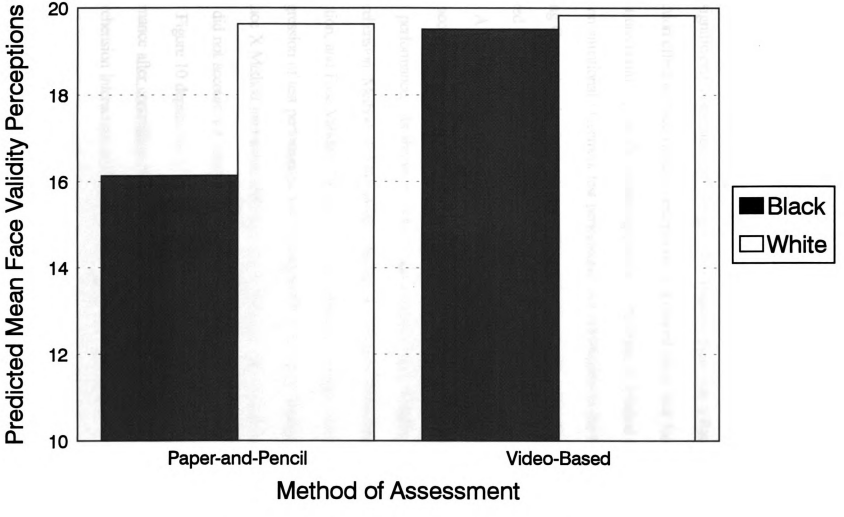

Figure 9. Race X Method Interaction on
Face Validity Perceptions of Situational Judgement Test.

Relationships between Race, Reading Comprehension, Method of Assessment, Face Validity Perceptions, and Situational Judgement Test Performance

Face validity perceptions and performance on the situational judgement test were significantly correlated, $r = .33$, $p < .05$. Because there was a Race X Method interaction effect on face validity perceptions, it appeared likely that face validity perceptions could explain the remaining portion of the Race X Method interaction effect on situational judgement test performance not attributable to the Method X Reading Comprehension interaction on test performance (see results for H3). It should be noted that this result was not hypothesized.

A hierarchical regression was performed to examine if face validity perceptions could account for the remaining unaccounted portion of the Race X Method interaction on test performance. As shown in Table 7, the variables Race, Reading Comprehension, Method of Assessment, Method X Reading Comprehension interaction, and Face Validity Perceptions were entered as a single block in step 1 of the regression of test performance and accounted for 22% of the variance, $p < .05$. The Race X Method interaction term was then entered in step 2 of the regression, which did not account for unique variance, $\Delta R^2 = .008$, n.s.

Figure 10 depicts the plot of the Race X Method interaction on test performance after controlling for the effects of both Method X Reading Comprehension interaction and Face Validity Perceptions. Compared to Figures 5 and 7, Figure 10 shows that the Race X Method interaction disappeared after controlling for the effects of Method X Reading Comprehension and Face Validity Perceptions.

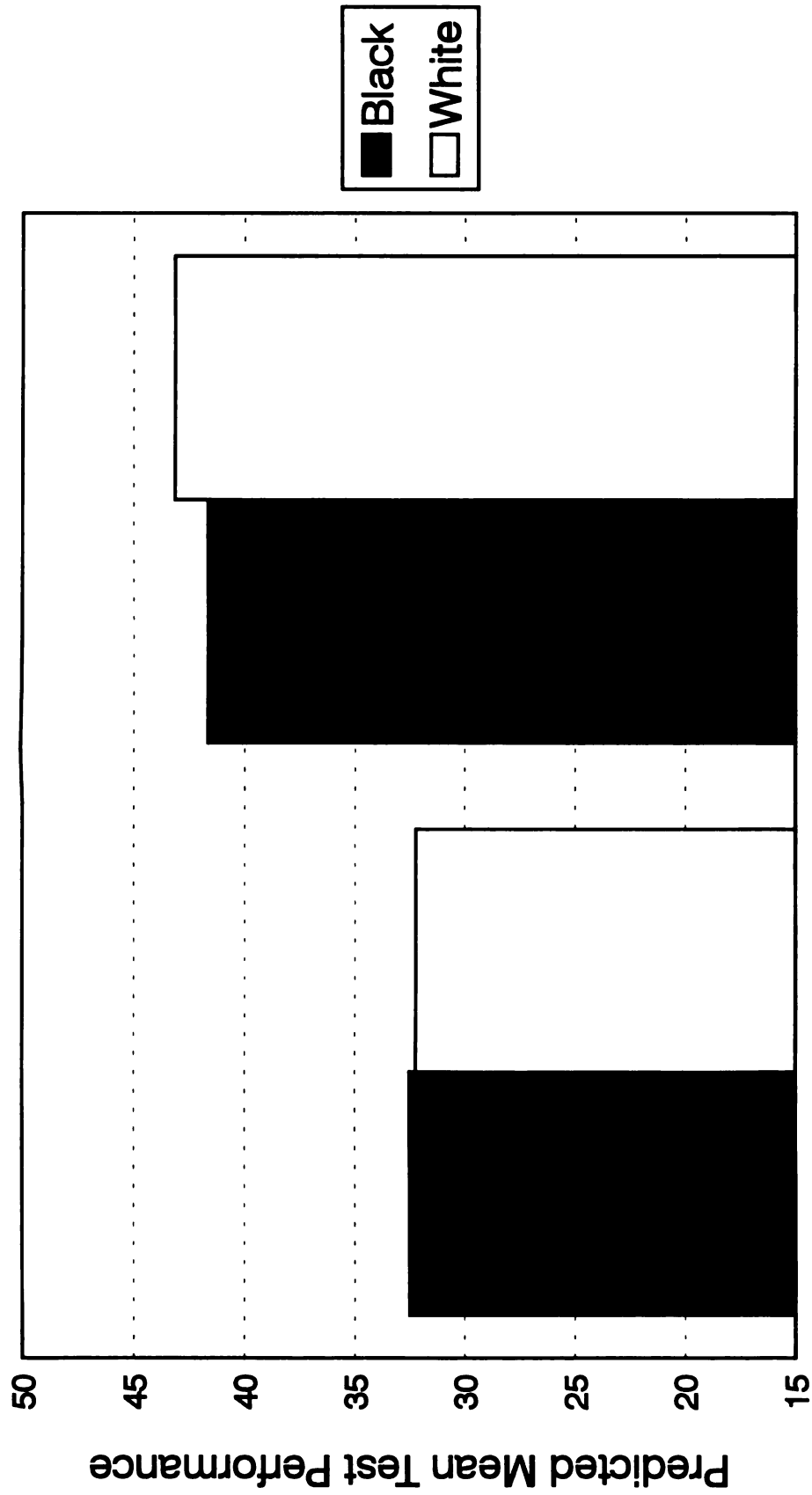

Figure 10. Race X Method Interaction on Situational Judgement Test Performance after Controlling for Effects of Method X Reading Comprehension Interaction and Face Validity Perceptions.

One implication of these results is that the use of a video-based method of item presentation might have had a "motivational" effect on Black examinees that affected their performance on the test. This idea will be discussed further below.

Figure 11 summarizes the relationships between Race, Reading Comprehension, Method of Assessment, Face Validity Perceptions, and Situational Test Performance.

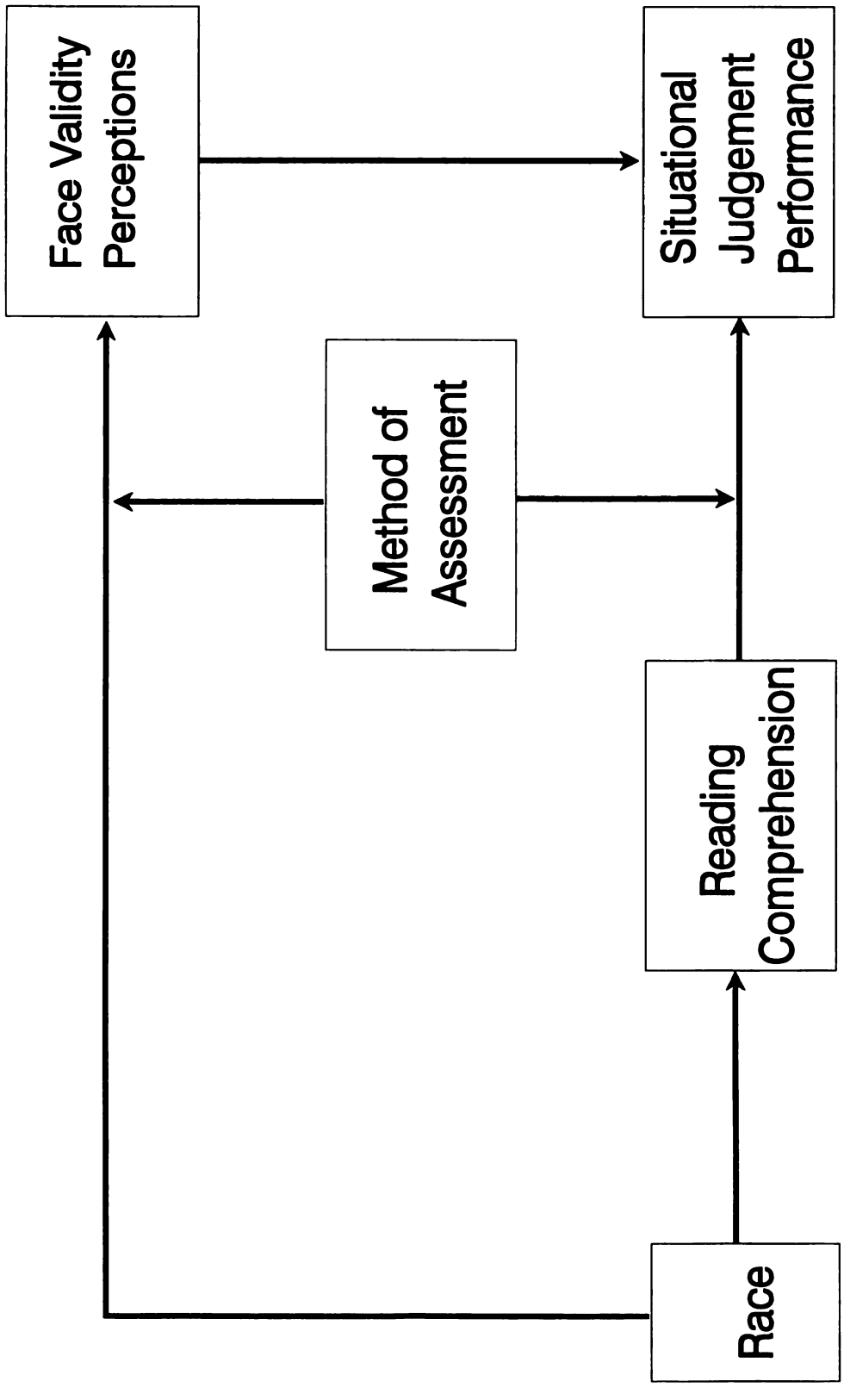

Figure 11. Relationships between Race, Reading Comprehension, Method of Assessment, Face Validity Perceptions, and Situational Test Performance.

DISCUSSION

The present study has established several theoretically and practically important effects relating race, reading comprehension, method of assessment, face validity perceptions, and performance on a situational judgement test. As predicted by H1, race and the method of assessment interact to affect situational judgement test performance such that the Black-White performance difference (favoring Whites) is substantially smaller in the video-based method of testing than in the paper-and-pencil method. As predicted by H2, the method of assessment also interacts with examinees' reading comprehension ability such that test performance positively correlates with reading comprehension ability in the paper-and-pencil method but performance and reading comprehension are nearly uncorrelated in the video-based method. The results for H3 supported the argument that this Method X Reading Comprehension interaction accounts for a substantial portion of the Race X Method interaction effect on test performance.

Another set of important results involved examinee reactions to the situational judgement test. As predicted by H5, face validity perceptions of the test are significantly higher when administered in the video-based method than when administered in the paper-and-pencil method. In addition, race and the method of assessment interact to affect face validity perceptions in a manner as predicted by H6. The difference in face validity perceptions reported by Blacks and Whites (with Whites giving higher face validity ratings) is greater in the paper-and-pencil method than in the video-based method. Finally, the results also suggest that face validity

perceptions may explain the remaining portion of the Race X Method interaction effect on situational judgement test performance not attributable to the Method X Reading Comprehension interaction on test performance.

The implications and contributions of the present study to the research on subgroup differences in test peformance and test reactions extend beyond the study of situational judgement tests. The issues revolve around the relationships between the method-content distinction and subgroup differences in test performance and test reactions. These issues will be discussed next in terms of conceptual, methodological, and practical implications.

Method-Content Distinction

A fundamental contribution of the present study is the emphasis on the distinction between test method and test content. By disconfounding method and content in the present study, subgroup differences due to method and subgroup differences due to content can be isolated. By holding test content constant, the Race X Method interaction effect on test performance obtained in the present study shows that two different methods of testing measuring the same job-relevant content may have differential adverse impact. In principle, adverse impact due solely to method of testing can be eliminated by using the method with lower adverse impact assuming that method is job-irrelevant.

Schmitt et al. (1996) argued that a significant amount of the Black-White difference in performance on paper-and-pencil tests might be due solely to the

reading/written requirements inherent in the method of testing and independent of the test content. As discussed earlier, Goldstein et al.'s (1993) attempt to show that the method of testing can affect differences in subgroup test performance has several problems. The method versus content distinction made in the present study enables an empirical test of Schmitt et al.'s (1996) argument. In addition, the inclusion of a standard reading comprehension test in the study allows a direct test of the notion of a Method X Reading Comprehension interaction effect on test performance. The present findings regarding H1, H2, and H3 also support the argument that race differences in test scores may be partly the result of differences in the reading requirements associated with the method of testing.

Test Reactions

The present study contributes to the recent research on examinee reactions toward selection procedures. The only study which attempted to examine the relationship between face validity and predictive validity perceptions is Smither et al. (1993). As discussed earlier in the paper, interpretations of the study's findings are problematic because perceptions measured were based on an examination consisting of a variety of selection procedures and the examinees used were applicants assessed for a variety of jobs ranging from entry-level to professional positions. The present study avoided these problems by using the job of a production worker as the frame of reference and examining the relationship between face validity and predictive validity perceptions separately for four different tests. As predicted by H4, face validity

perceptions and predictive validity perceptions are positively and strongly correlated. In addition, for each test, the correlation between the two types of perceptions was substantially lower than the internal consistency reliability estimates of the respective perception measures therefore providing evidence of discriminant validity for the two types of perceptions.

In the present study, the investigation of examinee test reactions is integrated into the broader selection framework by analyzing subgroup differences in test reactions and examining its relationship to adverse impact and method of testing. Previous studies which simply compared and described mean differences in attitudes or reactions across tests have been limited in increasing our understanding of test reactions due to the method-content confound across tests. The method-content distinction helps clarify the aspects of tests responsible for examinee reactions. The results of the present study show that without varying test item content, the method of testing per se can affect face validity perceptions, including subgroup differences in these perceptions.

A serendipitous finding (insofar as the results were not hypothesized) in the present study relates to the role of face validity perceptions in explaining the remaining portion of the Race X Method interaction effect on situational judgement test performance not attributable to the Method X Reading Comprehension interaction on test performance. Race and method of assessment interact to affect face validity perceptions which in turn affect test performance. In other words, subgroup differences in reading comprehension may account for a substantial portion of the

Black-White difference in test performance in a paper-and-pencil method of assessment. In addition, a nontrivial part of the adverse impact could be due to the fact that the paper-and-pencil method of assessment elicits lower face validity perceptions from Black examinees relative to White examinees. This lowered face validity may have a negative motivational and performance effect on Black examinees.

The present results regarding the relationships between race, method of assessment, face validity, and test performance contribute to the recent research on test reactions. Face validity perceptions constitute an important dimension of test reactions. Some researchers have argued that low face validity could result in biased or inaccurate test scores and reduce the operational validity of a selection procedure (e.g., Cascio, 1987; Robertson & Kandola, 1982; Smither, et al., 1993). Chan, Schmitt, DeShon, Clause, & Delbridge (under review) provided evidence that face validity perceptions affect test-taking motivation which in turn affects cognitive test performance. Chan et al. also found that the typical Black-White difference in test performance was partially mediated by differences in face validity perceptions and test-taking motivation. Arvey et al. (1990) argued that the traditional model of cognitive test performance as simply a function of ability plus error is probably incorrect and that researchers have tended to focus exclusively on the ability dimension and have ignored the effort dimension or motivational aspects of test performance. A similar argument may apply to performance on situational judgement tests. The present results suggest that the Black-White difference in performance on a paper-and-pencil situational judgement test could be decomposed into an ability

component (i.e., reading comprehension dfferences) and a motivational component (i.e., face validity perception differences). However, a difference between the situational judgement test and the traditional cognitive ability test is that in the former, the ability (i.e., reading comprehension) dimension is often not part of the construct space intended to be measured by the test and is therefore job-irrelevant.

Chan et al. argued that an important practical implication of their findings was that face validity of a test represents a practical means of reducing adverse impact of many traditional paper-and-pencil measures because it is possible to write test items that reflect a credible face valid relationship to the performance of jobs for which examinees are being assessed. The present study found that the manipulation of the method of test item presentation resulted in changes in face validity perceptions including changes in the size of the Black-White difference in these perceptions. It is plausible that these changes in perceptions in turn affected the Black-White difference in test performance. Whereas it is possible to affect face validity perceptions by writing credible items, the present findings suggest that simply changing the method of item presentation without changing item content may have substantial effects on subgroup differences in face validity perceptions and test performance.

Although the present results from the regression analyses are consistent with the idea that face validity perceptions affect test performance, it is also possible that test performance affects face validity perceptions. Chan et al. suggested that examinees' performance on a cognitive ability test may influence subsequent responses to face validity items. A self-serving mechanism may operate for reported face

validity such that there exists a tendency for examinees to attribute poor test performance to low face validity of the test. Poor performance on a test in which its content is perceived as unrelated to the content of the job is more self-serving than when test content is perceived as related to the content of the job. However, a self-serving bias explanation is a weaker argument in the case of performance on situational judgement tests than in the case of performance on traditional cognitive ability tests. This is because it is more difficult for an examinee to have knowledge or an estimate of his or her performance level on a situational judgement test compared to a cognitive ability test.

It is not the purpose of the present study to address the causal relationships between face validity perceptions and test performance. The present data relating face validity perceptions and test performance are correlational in nature and causal inferences are not possible. Future research should consider experimental designs for manipulating test reactions and examining if Black-White differences in test performance can be reduced by changes in test reactions.

Factorial Invariance of Test Responses across Assessment Methods

Although method and content are conceptually distinct, it is often difficult to separate the two empirically. The Goldstein et al. (1993) study discussed earlier in this paper illustrates the methodological difficulty in isolating the effects of method from the effects of test content and vice versa. The present study suggests that one way to tease out the two different effects is to examine a common set of test items

across different methods of testing. By holding test item content constant, the same intended constructs are presumably held constant across methods.

However, holding item content constant does not guarantee that the same constructs are measured across method groups. Measurement invariance of responses to the test items is critical and needs to be established. In the absence of established measurement invariance, there is no support for meaningful between-method comparisons of test scores. As demonstrated in the present study, measurement invariance can be tested using the multiple-group approach to confirmatory factor analysis. Ideally, the researcher should have a priori scales for the constructs of interest so that he or she can proceed to test for equality of relevant parameter estimates (e.g., factor loadings, factor covariances, error variances) across method groups in a theory-driven manner.

Another way to test if the same constructs are measured across different method groups by holding test items constant is through the assessment of nomological invariance. The idea is similar to the assessment of external parallelism in the classical psychometric development of test items. Given a set of external reference variables, some of which are expected (by some conceptual reasons) to be empirically related to the constructs measured on the test whereas others are not, we have evidence of factorial invariance of responses to the test across method groups if both groups exhibit the same patterns of correlations between test constructs and external variables. In the present study, nomological invariance of test responses across the two method groups could not be tested because of near-zero correlations

between situational judgement factors and the external reference personality variables. Therefore, the researcher should base the search and choice of external variables on solid theoretical grounds and relevant previous empirical literature. Of course, this is often not easy because it presupposes that the researcher has little difficulty in explicating the nature of the constructs of interest on the test examined which may not always be the case.

The mean differences obtained in the present study between racial subgroups and between methods indicate the presence of reading comprehension and some motivational difference associated with race and method. It should be noted that these mean differences reflect level differences on the situational judgement factors due to the effects of reading comprehension and motivational differences. Mean differences are consistent with factorial invariance of test responses across method groups (in terms of both measurement invariance and nomological invariance). The same construct can be measured in two groups though the groups may differ with respect to the level on the construct. Measurement invariance can coexist with mean differences because differences in factor means across method groups are independent of the equality of item-factor loadings, error variances, and factor covariances across method groups. Nomological invariance can coexist with mean differences because differences in factor means across method groups are independent of the equality of correlations between factors and external reference variables.

Measurement Errors and Effect Size Estimates

Another methodological issue concerns the need to correct effect size estimates (for subgroup differences) for attenuation due to unreliability. The majority of previous studies comparing adverse impact across selection procedures failed to report reliability estimates for the various measures or failed to correct effect size estimates for attenuation due to unreliability of measurement. With low reliabilities, true subgroup differences will not be detected. For studies reporting differential adverse impact across measures based on uncorrected effect size estimates, it is not clear if the results are due to true subgroup differences or simply an artifact of differential reliability in measurement. In the case of situational judgement tests, the difficulty is compounded because Cronbach's α, the most readily available reliability estimate, may not be an appropriate reliability index due to the multidimensional nature of these tests. Test-retest reliability is hard to obtain because it requires at least two separate administrations of the same test to the same examinees. Parallel form reliability is often not feasible because it requires the use of different item content which raises the issue of construct equivalence and complicates the interpretation of corrected estimates.

The present study suggests that one way to examine corrected effect size estimates for the multidimensional situational judgement test is to compute, for all examinees, factor scores for each situational judgement factor based on the factor loadings in the conceptually derived and empirically validated measurement model. Because the factors are latent variables free of measurement errors in the observed indicators, comparisons of method group and racial subgroup differences in factor

scores provide more accurate estimates (i.e., disattenuated for unreliability in measures) of the effect of method of assessment on test performance and adverse impact.

Limitations and Future Research

At least three limitations of the present findings should be noted. The first limitation concerns the generalizability of the findings relating to face validity perceptions. There are settings in which all examinees are likely to report that all tests are highly face valid. Examples of these settings include testing situations of actual job applicants or incumbents in which the stakes for successful test performance are high (e.g., assessment for hiring or promotion). It is very unlikely that an applicant taking a selection test for a job to which he or she desires to be hired will report low face validity on the test. In these high stake situations, self-presentation concerns may restrict reported face validity to high ratings when examinees perceive that test reactions may be used as inputs to individual situations. This is most likely to happen when examinees do not have confidence that face validity responses are anonymous. In such settings, restriction of range limits the effect size estimates associated with face validity perceptions. However, it should be noted that in many of these settings, the assessment of face validity is likely to have low construct validity. Future research on the face validity of different testing methods should be sensitive to the nature of the samples used and the setting of the test assessment situation. Theories and measures of social desirability and self-presentation concerns may be relevant in

certain high stake settings.

A second limitation concerns the nature of the constructs measured in the situational judgement test. Although the study addressed limitations in previous research by focusing on a priori situational judgement factors, correcting for measurement errors, and establishing factorial invariance of test responses across methods, more work needs to be done on construct validation. At this point, it is premature to use scores on individual situational judgement factors (at least those measured in this study) for any individual diagnostic or decision purpose. Future research should be explicit in the preoperational constitutive definitions of the relevant constructs in order to guide the development of appropriate measures (i.e., writing valid items). Finally, nomological invariance was not tested in the present study due to the inappropriate choice of external reference variables. In future research, factorial invariance of test responses across method groups in terms of both measurement invariance and nomoloigical invariance should be empirically established and not merely assumed.

The focus of the present study was not on test bias as defined by the Cleary model (Cleary, 1968). No criterion performance data were collected to examine differential prediction across racial subgroups and method groups. From a practical perspective, future research should examine potential relationships between differential prediction and method effects on subgroup differences in test performance and face validity perceptions (or other motivational variables). For example, consider the use of test scores on the paper-and-pencil version of the situational judgement test in the

present study as a predictor of job performance. If reading comprehension is job-irrelevant and uncorrelated with actual job performance, then using a common regression line based on the regression of job performance on situational judgement test scores would likely result in an over-prediction for White examinees and under-prediction of Black examinees. That is, test bias in the Cleary sense would occur.

Conclusion

The present study contributes to the sparse research on video testing in personnel selection and the research on situational judgement testing in particular. As mentioned early in the paper, the only published study reporting the adverse impact level of the video-based situational judgement test (Smiderle et al., 1994) did not correct for unreliability of measurement. The present study reports corrected estimates and isolates the method and content sources of subgroup differences in video testing. With the increasing popularity of video testing, clarifying the nature of its associated adverse impact levels becomes important from a legal and socio-political perspective.

With the exception of relatively higher costs in test development due to video production, the video-based method shares the same practical benefits with the paper-and-pencil format of the situational judgement test including the scale of testing which allows a large number of examinees in one session. Moreover, the video-based method is more realistic and concrete than the paper-and-pencil method. The method also elicits less adverse impact, more favorable face validity reactions in general and less subgroup differences in these reactions in particular. In addition, the video-based

method is generally less expensive than such high fidelity simulations as work samples and assessment centers. Hence, from a practical perspective, it is worthwhile to invest more research efforts in video-based testing and compare the method with the traditional paper-and-pencil method of assessment for the same test. The method-content distinction made in the present study provides a conceptual and methodological basis for formulating future study designs.

REFERENCES

Alwin, D.F., & Jackson, D.J. (1981). Application of simultaneous factor analysis to issues of factor invariance. In D.J. Jackson & E.F.Borgatta (Eds.), Factor analysis and measurement in sociological research (pp.249-279). Beverly Hills, CA: Sage.

Arvey, R.D., & Sackett, P.R. (1993). Fairness in selection: Current developments and perspectives. In N.Schmitt & W.C.Borman (Eds.), Personnel selection in organizations (pp.171-202). San Francisco: Jossey-Bass Publishers.

Arvey, R.D., Strickland, W., Drauden, G., & Martin, C. (1990). A Motivational components of test taking. Personnel Psychology, 43, 695-716.

Asher, J.J., &, Sciarrino, J.A. (1974). Realistic work sample tests: A review. Personnel Psychology, 27, 519-533.

Bentler, P.M. (1990). Comparative fit indexes in structural models. Psychological Bulletin, 107, 238-246.

Bentler, P.M. & Bonett, D.G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. Psychological Bulletin, 88, 588-606.

Bernardin, H.J. (1984). An analysis of black-white differences in job performance. Paper presented at the 44th Annual Meeting of the Academy of Management, Boston.

Bobko, P., & Bartlett, C.J. (1978). Subgroup validities: Differential definitions and differential prediction. Journal of Applied Psychology, 63, 12-14.

Bollen, K.A. (1989). Structural equations with latent variables. New York: Wiley.

Boudreau, J. (1983). Economic considerations in estimating the utility of human resource productivity improvement programs. Personnel Psychology, 36, 551-576.

Brown,J., Bennet, J., & Hanna, G. (1993). The Nelson-Denny Reading Test. Lombard, IL: Riverside.

Bruce, M.M., & Learner, D.B. (1958). A supervisory practices test. Personnel Psychology, 11, 207-216.

Brugnoli, G.A., Campion, J.E., & Basen, J.A. (1979). Racial bias in the use of work samples for personnel selection. Journal of Applied Psychology, 64, 119-123.

Cascio, W.F. (1982). Costing human resources: The financial impact of behavior in organizations. Boston, MA:Kent.

Cascio, W.F., & Phillips, N. (1979). Performance testing: A rose among thorns? Personnel Psychology, 32, 751-766.

Casio, W.F. (1987). Applied psychology in personnel management (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall, Inc.

Chan, D. (in press). Criterion and construct validation of an assessment center. Journal of Occupational and Organizational Psychology.

Chan, D., Schmitt, N., DeShon, R.P., Clause, S.C., Delbridge, K. Reactions to cognitive ability tests: The relationships between race, test performance, face validity, and test-taking motivation. Manuscript under review.

Cleary, T.A. (1968). Test bias: prediction of grades of negro and white students in integrated colleges. Journal of Educational Measurement, 5, 115-124.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Erlbaum.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. 2nd edition. Hillsdale, NJ: Erlbaum.

Cohen, J., & Cohen, P. (1983). Applied MRC analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Costa, P.T., Jr., & McRae, R.R. (1985). The NEO Personality Inventory manual. Odessa, Florida: Psychological Assessment Resources, Inc.

Costa, P.T., Jr., & McRae, R.R. (1992). NEO PI-R: Professional Manual. Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI). Odessa, Florida: Psychological Assessment Resources, Inc.

Dalessio, A.T. (1994). Predicting insurance agent turnover using a video-based situational judgement test. Journal of Business and Psychology, 9(1), 23-32.

Digman, J.M. (1990). Personality structure: Emergence of the five factor model. Annual Review of Psychology, (Vol.41, Pp.417-460). Palo Alto, CA: Annual Reviews.

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are central issues. Psychological Bulletin, 95, 134-135.

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. Journal of Applied Psychology, 72, 19-29.

Dyer, P.J., Desmarais, L.B., Midkiff, K.R. (1993). Multimedia employment testing in IBM: Preliminary results from employees. Paper presented at the Eighth Annual Conference of the Society for Industrial and Organizational Psychology, San

Francisco.

File, Q.W., & Remmer, H.H. (1971). How Supervise? (Manual, Rev.ed.). NY: Psychological Corporation.

Gaugler, B.B., Rosenthal, D.B., Thornton, G.C., & Bentson, C. (1987). Meta-analysis of assessment center validity. Journal of Applied Psychology, 72 (3), 493-511.

Gilliland, S.W. (1993). The perceived fairness of selection systems: An organizational justice perspective. Academy of Management Review, 18, 694-734.

Gilliland, S.W. (1994). Effects of procedural and distributive justice on reactions to a selection system. Journal of Applied Psychology, 79, 691-701.

Goldstein, H.W., Braverman, E.P., & Chung, B. (1993). Method versus content: The effects of different testing methodologies on subgroup differences. Paper presented at the Eighth Annual Conference of the Society for Industrial and Organizational Psychology, Montreal, Canada.

Helms, J.E. (1992). Why is there no study of cultural equivalence in standardized cognitive ability testing? American Psychologist, 47, 1083-1101.

Herriot, P. (1989). Selection as a social process. In N.Smith & I.Robertson (Eds.), Advances in selection and assessment. NY: Wiley.

Huck, J.R., & Bray, D.W. (1976). Management assessment center evaluations and subsequent job performance of black and White females. Personnel Psychology, 29, 13-30.

Hunter. J.E., & Hunter, R.F. (1984). Validity and utility of alternative predictors of job performance. Psychological Bulletin, 96, 72-98.

Hunter, J.E., Schmidt, F.L., & Hunter.R.F. (1979). Differential validity of

employment tests by race: a comprehensive review and analysis. Psychological Bulletin, 86, 721-735.

Iles, P.A. & Robertson, I.T. (1989). The impact of personnel selection procedures on candidates. In P.Herriot (Ed.), Assessment and selection in organizations (pp.257-271). Chichester, England,: Wiley.

James, L.R. & James, L.A. (1989). Causal modeling in organizational research. In C.L., Copper and I.Robertson (eds), International Review of Industrial and Organizational Psychology, 1989 (pp.371-404). Chichester, UK: John Wiley.

Jensen, A.R. (1980). Bias in mental testing. NY: Free Press.

Joreskog, K. & Sorbom, D. (1986). LISREL VI: Analysis of Linear Structural Relationships by Maximum Likelihood and Least Square Methods. Mooresville, IN: Scientific Software, Inc.

Joreskog, K. & Sorbom, D. (1989). LISREL 7: A guide to the program and applications (2nd ed.). Chicago, SPSS.

Joreskog, K. & Sorbom, D. (1993). Lisrel 8: Structural equation modeling with the SIMPLIS command language. Chicago: Scientific Software.

Kelloway, E.K. (1996). Common practices in structural equation modeling. In C.L., Copper and I.Robertson (eds), International Review of Industrial and Organizational Psychology, 1996 (pp.141-180). Chichester, UK: John Wiley.

Klimoski, R.J., & Brickner, M. (1987). Why do assessment centers work? The puzzle of assessment center validity. Personnel Psychology, 40, 243-260.

Latham, G.P., & Saari, L.M. (1984). Do people do what they say? Further studies on the situational interview. Journal of Applied Psychology, 69, 569-573.

Latham, G.P., Saari, L.M., Pursell, E.D., & Campion, M.A. (1980). The

situational interview. Journal of Applied Psychology, 65, 422-427.

Linn, R.L. (1978). Single group validity, differential validity, and differential prediction. Journal of Applied Psychology, 63, 507-514.

Loehlin, J.C., Lindzey, G., & Spuhler, J.M. (1975). Race differences in intelligence. San Francisco: Freeman.

Macan, T.H., Avedon, M.J., Paese, M., & Smith, D.E. (1994). The effects of applicants' reactions to cognitive ability tests and an assessment center. Personnel Psychology, 47, 715-738.

Matthews, D.B. (1991). Learning styles research: Implications for increasing students in teacher education programs. Journal of Instructional Psychology, 18, 228-236.

McDaniel, M.A., Whetzel, D.L., Schmidt, F.L., & Maurer, S.D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. Journal of Applied Psychology, 79, 599-616.

McHenry, J.J., & Schmitt, N. (1994). Multimedia testing. In M.G.Rumsey, C.B.Walker, & J.H.Harris (Eds.), Personnel selection and classification. NJ: Hillsdale.

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., & Tippins, N. (1993). Further studies of the low-fidelity simulation in the form of a situational inventory. Journal of Occupational and Organizational Psychology, 66, 337-344.

Premack, S.L., & Wanous, J.P. (1985). A meta-analysis of realistic job preview experiments. Journal of Applied Psychology, 70, 706-719.

Pulakos, E.D., Schmitt,N., & Keenan, P.A. (1994). Validation and Implementation of the FBI special agent entry-level selection system. (HumRRO Final Report FR-PRD-94-20). Alexandria, VA: Human Resources Research Organization.

Reily, R.R., & Chao, G.T. (1982). Validity and fairness of some alternative employee selection procedures. Personnel Psychology, 35, 1-62.

Robertson, I.T., & Smith, M. (1989). Personnel selection methods. In M.Smith & I.T.Robertson (Eds.), Advances in selection and assessment (pp.89-112). NY:Wiley.

Sackett, P.R., & Dreher, G.F. (1982). Constructs and assessment center dimensions: Some troubling empirical findings. Journal of Applied Psychology, 67, 401-410.

Sackett, P.R., & Harris, M.M. (1983). A further examination of the constructs underlying assessment center ratings. Paper presented at the American Psychological Association Convention, Anaheim, CA.

Schmidt, F.L., & Hunter, J.E. (1981). Employment testing: Old theories and new research findings. American Psychologist, 36, 1128-1137.

Schmidt, F.L., Greenthal, A.L., Hunter, J.E., Berner, J.G., & Seaton, F.W. (1977). Job samples vs. paper-and-pencil trade and technical tests: Adverse impact and examinee attitudes. Personnel Psychology, 30, 187-197.

Schmitt, N. (1993). Group composition, gender, and race effects on assessment center ratings. In H.Schuler, J.L. Farr, & M.Smith (Eds.), Personnel selection and assessment: Individual and organizational perspectives. NJ: Hillsdale.

Schmitt, N., Clause, C.S., & Pulakos, E.D. (1996). Subgroup differences associated with different measures of some common job relevant constructs. In C.L.

Cooper & I.T. Robertson (Eds.), International Review of Industrial and Organizational Psychology. NY: Wiley.

Schmitt, N., & Gilliland, S.W. (1992). Beyond differential prediction: Fairness in selection. In D. Saunders (Ed.), Human rights and employment: Interdisciplinary perspectives (Vol.1, pp.21-46). Greenwich, CT: JAI Press.

Schmitt, N., Gilliland, S.W., Landis, R.S., & Devine, D. (1993). Computer-based testing applied to selection of secretarial applicants. Personnel Psychology, 46, 149-165.

Schmitt, N., Gooding, R.Z., Noe, R.A., & Kirsch, M.P. (1984). Meta-analyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. Personnel Psychology, 37, 407-422.

Schmitt, N., & Noe, R.A. (1986). Personnel selection and equal employment opportunity. In C.L. Cooper & I.T. Robertson (Eds.), International Review of Industrial and Organizational Psychology. NY: Wiley.

Schneider, J., & Schmitt, N. (1992). An exercise design approach to understanding assessment center dimension and exercise constructs. Journal of Applied Psychology, 77, 32-41.

Schuler, H. (1993).Social validity of selection situations: A concept and some empirical results. In H.Schuler, J.L. Farr, & M.Smith (Eds.), Personnel selection and assessment: Individual and organizational perspectives. NJ: Hillsdale.

Schuler, H., & Fruhner, R. (1993). Effects of assessment center participation on self esteem and on evaluation of the selection situation. In H.Schuler, J.L. Farr, & M.Smith (Eds.), Personnel selection and assessment: Individual and organizational perspectives. NJ: Hillsdale.

Scott, R. (1987). Gender and race achievement profiles of Black and White third-grade students. Journal of Psychology, 121, 629-634.

Smith, M. & George, D. (1992). Selection methods. In C.L. Cooper & I.T. Robertson (Eds.), International Review of Industrial and Organizational Psychology. NY: Wiley.

Smiderle, D., Perry, B.A., & Cronshaw, S.F. (1994). Evaluation of video-based assessment in transit operator selection. Journal of Business and Psychology, 9(1), 3-22.

Smither, J.W., Reilly, R.R., Millsap, R.E., Pearlman, K., & Stoffey, R.W. (1993). Applicant reactions to selection procedures. Personnel Psychology, 46, 49-76.

Society for Industrial and Organizational Psychology. (1987). Principles for the validation and use of personnel selection procedures. (Third Edition). College Park, MD: Author.??

Sorbom, ,D. (1974). A general method for studying differences in factor means and factor structures between groups. British Journal of Mathematical and Statistical Psychology, 27, 229-239.

Steiger, J.H. (1990). Structural model evaluation and modification: An interval estimation approach. Multivariate Behavioral Research, 25, 173-180.

Tenopyr, M.L. (1969). The comparative validity of selected leadership skills relative to success in production management. Personnel Psychology, 22, 77-85.

Turnage, J.J., & Muchinsky, P.M. (1984). A comparison of the predictive validity of assessment center evaluations versus traditional measures in forecasting supervisory job performance: Interpretive implications of criterion distortion for the assessment center. Journal of Applied Psychology, 69, 595-602.

Uniform Guidelines on Employee Selection (1978). Federal Register, 43, 38290-38315.

Weekley, J.A., & Gier, J.A. (1987). Reliability and validity of the situational interview for a sales position. Journal of Applied Psychology, 72, 484-487.

Wernimont, P.R., & Campbell, J.P. (1968). Signs, samples, and criteria. Journal of Applied Psychology, 52, 372-376.

Wigdor, A.K., & Green, B.F., Jr. (1991). Performance assessment for the workplace. Washington, DC: National Academy Press.

Wilson Learning (1990). Validation report for the Teller Assessment Program (TAP). Longwood, FL: Wilson Learning.

Wonderlic, E.F., and Assoc. (1984). Wonderlic personnel test manual. Northfield, IL: E.F. Wonderlic & Associates, Inc.

APPENDICES

APPENDIX A

APPENDIX A

Series of A Priori Power Analyses

For each of the following power analyses, the desired power was fixed at .80 and α was fixed at .05. Expected effect sizes were construed as "small" effect sizes (Cohen, 1988).

H1:

H1 tests the unique variance accounted for by the Race X Method term over and above the set of control variables consisting of Race and Method (Set A). A small ΔR^2 of .03 was arbitrarily expected. The expected R^2 for the entire set of predictors (Set A + Race X Method term) was arbitrary fixed at a conservative value of .10. Using Cohen and Cohen's (1983) formula for effect size f^2, we have

$$f^2 = \Delta R^2/(1 - R^2)$$
$$= .03/(1 - .10)$$
$$= .033$$

According to Cohen (1988), a f^2 value of .033 is construed as a small effect size.

Cohen & Cohen 's (1983) formula for required sample size n* is as follows:

$$n^* = (L/f^2) + k + 1$$

k refers to df for unique source of variance. We have k = 1. From the table of L values in Cohen & Cohen (1983), we have L = 7.85. Therefore, we have

$$n^* = (7.85/.033) + 1 + 1$$
$$= 239.8$$

H2:

H2 tests the unique variance accounted for by the Method X Reading Comprehension term over and above the set of control variables consisting of Method and Reading Comprehension (Set A). The same assumptions as H1 were made which resulted in the same sample size requirement (n* = 239.8).

APPENDIX A

H3:

H3 tests the unique variance accounted for by the Race X Method term over and above the set of control variables consisting of Race, Reading Comprehension, Method, and Reading Comprehension X Method (Set A). The same assumptions were made as H1 except that R^2 was fixed at a higher (but nevertheless conservative) value of .15 because of the larger number of variables in Set A. Using the same formulae in H1 resulted in a required sample size of 226.3.

H4:

H4 tests the significance of the Pearson correlation coefficient between Face Validity Perceptions and Predictive Validity Perceptions. A small effect size of $r = .20$ was arbitrarily expected. Based on Cohen's (1988) tables for sample size requirements for correlation coefficients, a desired power of .80 at $\alpha = .05$ indicated that a sample size of 194 was required.

H5:

H5 tests the difference in mean face validity perceptions between the paper-and-pencil method and the video-based method. A conservative d value of .30 was arbitrarily expected. Based on Cohen's (1988) tables for sample size requirements for t tests between means, a desired power of .80 at $\alpha = .05$ indicated that a sample size of 175 was required.

H6:

H6 tests the unique variance accounted for by the Race X Method term over and above the set of control variables consisting of Race and Method (Set A). The same assumptions as H1 were made for ΔR^2 and R^2 for Set A. Using the same formulae in H1 resulted in a required sample size of 239.8.

APPENDIX B

APPENDIX B

EXAMPLE OF A PAPER-AND-PENCIL VIGNETTE

The following is an example of a written vignette on the test booklet and some possible responses on the answer booklet in the paper-and-pencil version of the situational judgement test.

Example of a written vignette on the test booklet.

SITUATION 1

Jerry and Dennis are discussing how they should go about checking the machinery in the building. Jerry told Dennis that they should start at the West end of the building and work their way East so that the more important machinery will be taken care of first. Dennis disagreed as he thinks that since the East end is on break right now, it would be much faster to start East and work their way West. Jerry said that he has never seen anyone doing it that way and besides, the machinery at the West end is more critical. Dennis continues to disagree and thinks that they should start at the East end.

Jerry can respond in a number of ways. For each possible response described in the answer booklet, indicate its effectiveness on the rating scale provided.

APPENDIX B

Examples of possible responses on the answer booklet.

After you have read SITUATION 1, rate the effectiveness of responses below from Jerry's perspective.

1. Ask your supervisor to decide which method is better.

2. Convince Dennis that your method is best.

3. Agree to use Dennis' method.

4. Split the work in half. Each of you use your own method.

5. Tell Dennis that you will use his method for a while, but you will switch if it looks like your method is best.

6. Tell Dennis that he needs to listen carefully to your ideas.

7. Compromise. Use Dennis' method today and your method next time.

8. Demand that Dennis use your method.

For each possible response, the following rating scale is provided.

VERY INEFFECTIVE	INEFFECTIVE	SOMEWHAT INEFFECTIVE	SOMEWHAT EFFECTIVE	EFFECTIVE	VERY EFFECTIVE
_____	_____	_____	_____	_____	_____

APPENDIX C

APPENDIX C

TEST REACTIONS QUESTIONNAIRE

QUESTIONNAIRE ON THE TEST THAT YOU HAVE JUST COMPLETED

Consider the job of a production worker which requires working in team-based situations. To do the job well, the worker has to be technically competent and also be able to relate to other persons effectively. For such a job, indicate how much you agree or disagree with the following statements about the test that you have just completed by circling the appropriate number on the rating scale provided.

1. I did not understand what the test had to do with the job.

____1_____2_____3_____4_____5____

STRONGLY DISAGREE	DISAGREE	NEITHER AGREE NOR DISAGREE	AGREE	STRONGLY AGREE

2. I could not see any relationship between the test and what I think is required by the job tasks.

____1_____2_____3_____4_____5____

STRONGLY DISAGREE	DISAGREE	NEITHER AGREE NOR DISAGREE	AGREE	STRONGLY AGREE

3. It would be obvious to anyone that the test is related to the job tasks.

____1_____2_____3_____4_____5____

STRONGLY DISAGREE	DISAGREE	NEITHER AGREE NOR DISAGREE	AGREE	STRONGLY AGREE

4. The actual content of the test was clearly similar to the job tasks.

____1_____2_____3_____4_____5____

STRONGLY DISAGREE	DISAGREE	NEITHER AGREE NOR DISAGREE	AGREE	STRONGLY AGREE

5. There was no real connection between the test and the job tasks.

____1_____2_____3_____4_____5____

STRONGLY DISAGREE	DISAGREE	NEITHER AGREE NOR DISAGREE	AGREE	STRONGLY AGREE

6. Failing to pass the test clearly indicates that you can't do the job.

____1_____2_____3_____4_____5____

STRONGLY DISAGREE	DISAGREE	NEITHER AGREE NOR DISAGREE	AGREE	STRONGLY AGREE

7. I am confident that the test can predict how well an applicant will perform on the job.

____1_____2_____3_____4_____5____

STRONGLY DISAGREE	DISAGREE	NEITHER AGREE NOR DISAGREE	AGREE	STRONGLY AGREE

8. My performance on the test was a good indicator of my ability to do the job.

____1_____2_____3_____4_____5____

STRONGLY DISAGREE	DISAGREE	NEITHER AGREE NOR DISAGREE	AGREE	STRONGLY AGREE

9. Applicants who perform well on the test are more likely to perform well on the job than applicants who perform poorly.

____1_____2_____3_____4_____5____

STRONGLY DISAGREE	DISAGREE	NEITHER AGREE NOR DISAGREE	AGREE	STRONGLY AGREE

10. The employer can tell a lot about the applicant's ability to do the job from the results of the test.

____1_____2_____3_____4_____5____

STRONGLY DISAGREE	DISAGREE	NEITHER AGREE NOR DISAGREE	AGREE	STRONGLY AGREE

APPENDIX D

APPENDIX D

Table 8

Means, Standard Deviations, and Intercorrelations of Study Variables broken down by Race

Variables[a]	Means	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
White Examinees (N=128)																							
1. METHOD	.46	.50																					
2. SEX	.32	.47	10																				
3. PSJ	43.36	7.34	----	11																			
4. VSJ	43.55	8.27	----	01	----																		
5. READING	31.02	4.31	-01	13	31	09																	
6. CONSC	33.00	5.43	14	-06	-16	-14	13																
7. AGREE	30.80	6.08	01	-18	21	04	-13	25															

Table 8 continues

	Means	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
8. NEURO	22.01	7.88	-06	-19	-08	17	-22	-29	-25														
9. OPEN	31.24	5.55	05	00	02	07	-07	00	06	06													
10. EXTRAV	32.96	6.54	01	-20	-27	-23	-13	27	29	-16	11												
11. COGN	24.68	5.04	-07	10	14	41	42	00	-10	-22	12	-16											
12. FACE- PSJ	19.63	3.27	----	-26	10	----	-03	12	02	19	11	02	04										
13. PRED-PSJ	13.07	3.11	----	03	-02	----	16	17	02	-12	06	-10	05	35									
14. FACE-VSJ	19.83	2.95	----	10	----	20	16	15	10	-04	-05	-08	00	----	----								
15. PRED-VSJ	14.09	3.83	----	-03	----	05	16	11	08	06	-17	-03	06	----	----	21							
16. FACE-READ	14.26	3.59	08	11	-18	-14	03	16	-10	-04	-04	06	04	-14	-13	09	11	22					
17. PRED-READ	12.21	3.87	-07	12	10	-06	19	06	-03	-12	02	-13	21	-07	32	03	10	33	56				
18. FACE- NEO	16.62	3.25	00	-01	-12	-07	-04	08	-07	-11	-06	-03	-02	-09	11	06	17	12	30	27			
19. PRED-NEO	13.04	4.19	03	07	-07	17	09	-03	05	-08	00	-08	13	05	50	-15	00	27	18	49	49		

Table 8 continues

	Means	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
20. FACE-COGN	13.77	3.49	05	04	-12	-01	06	11	-05	-12	13	-01	10	-06	20	-15	06	37	41	43	26		
21. PRED-COGN	12.15	3.70	-09	09	-05	05	11	03	-12	-04	11	-13	19	-04	44	03	06	23	57	55	25	71	

Black Examinees (N=113)

1. METHOD	.55	.50							
2. SEX	.41	.49	-01						
3. PSJ	35.58	7.56	----	-32					
4. VSJ	41.92	6.88	----	-15	----				
5. READING	27.03	6.39	-19	-11	34	-08			
6. CONSC	31.31	6.52	-08	-01	01	24	05		
7. AGREE	29.89	6.97	-16	-25	39	26	21	22	

Table 8 continues

	Means	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
8. NEURO	22.28	8.15	16	-22	-19	-05	-21	-31	-15														
9. OPEN	28.76	5.92	-08	-04	25	00	17	-08	09	05													
10. EXTRAV	31.26	5.71	-15	-25	30	-03	06	15	30	-17	11												
11. COGN	21.50	5.50	-01	-03	24	00	37	-13	00	14	06	03											
12. FACE- PSJ	16.13	4.59	----	-27	34	----	29	04	-01	00	11	02	39										
13. PRED-PSJ	13.21	3.31	----	-06	10	----	13	10	-05	-04	04	04	07	29									
14. FACE-VSJ	19.51	3.00	----	-22	----	07	06	07	23	20	-12	-01	14	----	----								
15. PRED-VSJ	14.08	3.16	----	-21	----	31	00	04	39	07	-04	-09	01	----	----	30							
16. FACE-READ	12.55	3.36	08	-04	06	24	-09	08	-04	02	01	09	-01	-03	26	11	-16						
17. PRED-READ	11.33	3.26	18	01	-01	46	-12	09	-04	00	-07	00	04	12	33	18	15	54					
18. FACE- NEO	16.59	3.17	-16	-14	09	00	07	17	21	-15	00	28	-12	-02	18	36	15	15	11				
19. PRED-NEO	13.44	3.61	03	-15	03	38	05	22	19	-11	00	22	01	09	34	26	51	05	32	39			

Table 8 continues

	Means	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
20. FACE-COGN	12.24	3.05	09	07	-05	14	-02	05	-16	00	-07	-21	-11	-08	15	06	-03	56	46	-20	-06		
21. PRED-COGN	10.89	3.12	11	05	04	26	-06	01	-01	03	-12	-15	-02	-07	40	16	23	37	51	-07	39	65	

[a] Variables abbreviated: METHOD=Method of assessment in situational judgement test; SEX=Sex of examinee; PSJ=Paper-and-Pencil Situational Judgement; VSJ=Video-based Situational Judgement; READ=Nelson-Denny Reading Comprehension; NEO= NEO-FFI Personality; CONSC=Conscientiousness; AGREE=Agreeableness; NEURO=Neuroticism; OPEN=Openness; EXTRAV=Extraversion; FACE=Face Validity; PRED=Perceived Predictive Validity.

Note. METHOD and SEX are dummy-coded (Video-based=0, Paper-and-Pencil=1; Female=0, Male=1).

APPENDIX E

APPENDIX E

Observed Covariance Matrices of Indicators for Situational Judgement Factors (con = Conflict; qua = Quality; com = Committment; emp = Empathy)

Group 1: Paper-and-Pencil Method

	con1	con2	con3	qua1	qua2	qua3
con1	3.96					
con2	0.57	1.74				
con3	1.17	0.67	4.00			
qua1	0.12	0.20	0.17	2.07		
qua2	0.01	0.24	-0.14	0.16	1.28	
qua3	-0.06	0.44	-0.23	0.32	0.23	1.46
com1	0.61	0.31	0.52	0.10	-0.03	-0.03
com2	0.13	-0.10	0.47	-0.02	0.14	0.03
com3	0.24	0.28	-0.04	0.42	0.14	0.15
emp1	0.33	0.71	0.74	0.25	0.19	0.46
emp2	0.14	0.76	-0.18	0.18	0.02	0.36
emp3	- -	0.47	0.44	0.57	0.07	0.30

	com1	com2	com3	emp1	emp2	emp3
com1	3.02					
com2	0.76	2.77				
com3	0.94	0.54	2.28			
emp1	0.69	-0.22	0.09	2.86		
emp2	0.56	-0.20	0.50	0.83	2.36	
emp3	0.43	0.10	0.05	0.94	1.06	2.47

Group 2: Video-Based Method

	con1	con2	con3	qua1	qua2	qua3
con1	4.26					
con2	0.41	1.74				
con3	0.73	0.77	3.43			
qua1	-0.15	0.62	0.22	2.61		
qua2	0.21	0.34	0.14	0.40	1.20	
qua3	-0.07	0.20	-0.15	0.70	0.15	1.38
com1	0.57	0.25	0.22	-0.23	0.09	-0.02
com2	0.14	0.16	-0.12	0.26	0.06	0.08
com3	-0.02	0.23	0.07	0.26	0.12	0.23
emp1	0.12	0.59	0.15	0.61	0.10	0.17
emp2	0.02	0.08	-0.17	0.17	-0.04	0.11
emp3	0.54	0.28	0.03	0.54	0.28	0.36

	com1	com2	com3	emp1	emp2	emp3
com1	2.16					
com2	0.25	2.68				
com3	0.66	0.90	2.60			
emp1	0.47	0.07	0.20	2.63		
emp2	0.21	- -	0.38	0.45	1.97	
emp3	0.25	-0.22	0.21	0.25	0.24	2.73

APPENDIX F

APPENDIX F

Observed Covariance Matrices of Indicators for Situational Judgement Factors and Personality Factors (con = Conflict; qua = Quality; Com = Committment; emp = Empathy; neuro = Neuroticism; extra = Extraversion; open = Openness; agree = Agreeableness; consc = Conscientiousness)

Group 1: Paper-and-Pencil Method

	con1	con2	con3	qua1	qua2	qua3
con1	1.00					
con2	0.25	1.00				
con3	0.32	0.26	1.00			
qua1	0.04	0.12	0.06	1.00		
qua2	0.00	0.17	-0.07	0.10	1.00	
qua3	-0.04	0.29	-0.09	0.19	0.18	1.00
com1	0.20	0.15	0.14	0.04	-0.01	-0.02
com2	0.04	-0.02	0.12	0.00	0.10	0.02
com3	0.07	0.14	-0.02	0.20	0.10	0.10
emp1	0.10	0.35	0.22	0.12	0.11	0.25
emp2	0.06	0.40	-0.05	0.10	0.03	0.21
emp3	0.00	0.24	0.16	0.28	0.04	0.19
neuro	-0.01	-0.03	-0.10	-0.17	-0.12	-0.15
extra	0.15	0.13	0.19	0.05	0.03	0.04
open	0.17	0.16	0.19	0.13	-0.08	0.08
agree	0.28	0.22	0.20	0.07	0.06	0.09
consc	0.18	-0.16	0.10	0.09	0.06	0.06

	com1	com2	com3	emp1	emp2	emp3
com1	1.00					
com2	0.29	1.00				
com3	0.37	0.26	1.00			
emp1	0.25	-0.09	0.04	1.00		
emp2	0.23	-0.08	0.25	0.34	1.00	
emp3	0.17	0.03	0.02	0.38	0.47	1.00
neuro	-0.11	-0.05	-0.10	-0.10	-0.10	-0.07
extra	0.02	-0.08	-0.04	0.03	0.03	0.06
open	0.21	0.08	-0.03	0.01	0.09	0.16
agree	0.21	0.07	0.12	0.21	0.14	0.03
consc	0.03	0.05	0.14	-0.09	-0.16	-0.11

	neuro	extra	open	agree	consc
neuro	66.97				
extra	-6.27	35.73			
open	2.85	7.89	35.24		
agree	-10.62	13.54	9.67	44.87	
consc	-14.88	10.10	-0.02	8.05	40.42

APPENDIX F

Appendix F continued.

Group 2: Video-Based Method

	con1	con2	con3	qua1	qua2	qua3
con1	1.00					
con2	0.16	1.00				
con3	0.18	0.32	1.00			
qua1	-0.04	0.31	0.09	1.00		
qua2	0.11	0.24	0.08	0.21	1.00	
qua3	-0.03	0.13	-0.07	0.37	0.12	1.00
com1	0.21	0.13	0.08	-0.09	0.06	-0.01
com2	0.05	0.10	-0.05	0.08	0.02	0.03
com3	0.02	0.12	0.03	0.08	0.07	0.12
emp1	0.03	0.28	0.06	0.25	0.05	0.10
emp2	-0.01	0.08	-0.05	0.07	-0.01	0.07
emp3	0.16	0.13	0.01	0.21	0.16	0.19
neuro	0.05	0.14	0.03	0.10	0.13	-0.11
extra	-0.04	-0.26	-0.16	-0.22	-0.10	-0.15
open	-0.06	0.11	0.23	-0.04	-0.15	-0.05
agree	0.04	-0.03	0.11	0.07	0.14	0.08
consc	-0.02	-0.01	-0.08	-0.05	0.09	0.11

	com1	com2	com3	emp1	emp2	emp3
com1	1.00					
com2	0.10	1.00				
com3	0.31	0.35	1.00			
emp1	0.21	0.03	0.08	1.00		
emp2	0.10	0.02	0.18	0.21	1.00	
emp3	0.11	-0.07	0.08	0.10	0.12	1.00
neuro	0.03	0.09	0.17	-0.01	-0.06	0.11
extra	0.11	0.13	0.01	-0.08	-0.06	-0.04
open	0.22	-0.06	0.03	0.09	-0.01	-0.11
agree	0.16	0.06	0.01	0.10	-0.04	-0.03
consc	0.06	0.10	-0.04	0.05	-0.08	-0.09

	neuro	extra	open	agree	consc
neuro	61.11				
extra	-9.62	41.44			
open	1.98	1.81	33.36		
agree	-9.70	10.44	-3.07	39.73	
consc	-14.24	6.98	-0.79	10.71	32.11