Frame-of-Reference Effects on Personality Scale Scores and Criterion-Related Validity

Mark J. Schmit
University of Florida

Ann Marie Ryan
Bowling Green State University

Sandra L. Stierwalt
Kansas State University

Amy B. Powell
Psychological Assessment Resources, Inc.

Increased use of personality inventories in employee selection has led to concerns regarding factors that influence the validity of such measures. A series of studies was conducted to examine the influence of frame of reference on responses to a personality inventory. Study 1 involved both within-subject and between-groups designs to assess the effects of testing situation (general instructions vs. applicant instructions) and item type (work specific vs. noncontextual) on responses to the NEO Five-Factor Inventory (P. T. Costa & R. R. McCrae, 1989). Results indicated that a work-related testing context and work-related items led to more positive responses. A second study found differences in the validity of a measure of conscientiousness, depending on the frame of reference of respondents. Specifically, context-specific items were found to have greater validity. Implications for personnel selection are discussed.
NEO Personality Inventory (NEO-PI–R; Costa & McCrae, 1992), a widely used personality inventory: “I try to be courteous to everyone I meet.” The applicant indicates the extent to which he or she agrees or disagrees with the statement. One applicant may consider only work experiences when making a decision on an item response, whereas another applicant may take all aspects of his or her life into consideration. Those applicants who base their answers on work experiences may be providing information that is a better indicator of actual job performance than those applicants who use their overall life experiences as a reference to answer the questions. Providing the same frame of reference to all applicants (e.g., using items that specifically refer to behavior at work) may, therefore, improve the predictive validity of personality inventories.

There are theory and evidence from personality psychology that indicate frame-of-reference differences do exist and potentially cloud the predictability of behavior. Wright and Mischel (1987) suggested that some people may express quite reliable and stable patterns of behaviors, but these behaviors are contingent on situational conditions; they labeled these tendencies conditional dispositions. Although there is a connection between personality and behavior, the power to predict behavior on the basis of a personality trait may be limited to a fairly specific range of situations; Wright and Mischel provided initial support for this proposition (Shoda, Mischel, & Wright, 1989, 1993; Wright & Mischel, 1987). Support of Mischel’s propositions is evident in other recent personality theories (e.g., Thorne, 1989; Zuroff, 1986) and empirical evidence (Moskowitz, 1994; Roberts & Donahue, 1994).

A metatheoretical framework for personality theories recently developed by McCrae and Costa (in press) helps to put the possibility of conditional dispositions into perspective. They suggested that personality measures must ask questions about characteristic adaptations (i.e., the concrete manifestations of basic tendencies) in order to make inferences about underlying traits. Because characteristic adaptations are influenced by external factors (e.g., cultural norms or organizational norms), many characteristic adaptations may be situation specific, even though across both time and situations (in aggregation) the underlying trait is generally constant. Although an individual’s behavior may be limited to a specific range of behaviors (e.g., limited by role expectations either learned from past experience or imposed by the organization) in the critical work situations from which performance criteria are drawn, many other life situations are still available for the demonstration of the general underlying trait. All of the theories and the research findings mentioned here suggest that general personality inventories may say little about how an applicant would act in an actual work situation, because there is no specific frame of reference in which the respondent considers the given behaviors. What may be needed are items or instructions that provide a frame of reference related to the workplace.

A theory of personality-item response with a growing base of supporting evidence also offers support for using items with a work-related context. The self-presentation view of item responses suggests that individuals use personality items to present an image of themselves that may not be totally accurate but that is consistent with how they hope to be regarded by others (Hogan, 1982, 1991; Johnson, 1981; Leary & Kowalski, 1990; Schlenker & Weigold, 1992). Some people have clearer self-images of how they want to be viewed by others; some people are better than others at self-presentation (Hogan, 1991). Hogan (1991) suggested that “item endorsements reflect automatic and often nonconscious efforts on the part of test-takers to negotiate an identity with an anonymous interviewer” (p. 902). General personality-test items used in a personnel selection test may present error into the self-presentation process, because many of the items are difficult to connect with a specific work role or context. Some researchers have suggested that the use of face-valid items could reduce the validity of personality measures because of the potential for greater socially desirable responding, although recent research has shown that socially desirable responding may not be a major problem in personnel selection contexts (Hough et al., 1990). However, self-presentation theory suggests that greater face validity should increase test validity. Johnson (1981) noted that “the best strategy for designing a valid scale is not to make lying or misrepresentation difficult, but to make self-presentation as easy as possible” (p. 767). Putting personality items into a work-specific format would facilitate self-presentation.

The objective of our first study was to investigate the effects of the respondent’s frame of reference on personality-scale scores by altering personality-test items’ specificity (work specific vs. noncontextual) in different administration conditions (job-applicant instructions vs. general instructions). This study included both a between-groups substudy and a mixed-design substudy. More specifically, in the between-groups substudy, participants were randomly assigned to one of four conditions differing on instruction type and item specificity: (a) general instructions, noncontextual items, (b) general instructions, work-specific items, (c) applicant instructions, noncontextual items, and (d) applicant instructions, work-specific items. In the mixed-design substudy, participants were randomly assigned to either the general instructions group or the applicant instructions group (between), and they completed both the work-specific and noncontextual item forms of the personality test (repeated measures). Both types of designs were used to
compensate for the weaknesses of each (see Lautenschlager, 1986; Schwab, 1971).

In organizations, a general set of clearly defined expectations for behavior are likely to exist that suggest appropriate behavior or explicitly forbid inappropriate behavior (cf. Weiss & Adler, 1984). Thus, most individuals recognize to some extent that their behavior at work may be more constrained than their behavior in some other general life situations. Consequently, we hypothesized that mean test scores of individuals responding to work-specific personality items would have more positive scale scores than would individuals responding to more general context personality items. Similarly, because some individuals use work experience as a frame of reference when responding to items in a personnel selection context, whether the items are work specific or not, individuals in the applicant instructions condition were hypothesized to have more positive scale scores than those in the sincere response instructions condition.

The self-presentation theory of item response also suggests that an interaction between instruction type and item specificity should be evidenced. Because the applicant instructions present the clearest press for a specific social role to be presented and because work-specific items have clear connections with work roles, individuals in the applicant instructions and work-specific items conditions should score significantly higher than individuals in the general instructions and noncontextual items conditions, in which neither this press nor the connection to work roles is present.

Finally, the manipulation of item content could have psychometric consequences. Adding an "at work" tag to items could change a multifactor inventory to a one-factor inventory. If this were the case, then the new items and the associated scales would no longer be comparable to the test manual and to the relevant literature. The conceptual integrity of the scales would be called into question, and an argument might be made that personality was no longer being measured; in fact an argument might be made that the new items may approach a format more like biodata items than personality items. To assess the impact of item changes on psychometric properties of the measures, we conducted several confirmatory factor analyses that addressed both the multifactor versus one-factor question and the question of psychometric equivalence for the altered and unaltered forms of the personality measures used in this study.

Study 1
Method

Participants

Substudy 1. Students of an introductory psychology course participated for course credit. Data were collected from a total of 100 participants, with 25 participants randomly assigned to each of the four study conditions. Across groups, participants did not differ significantly in age, class rank, or the number of jobs previously held.

Substudy 2. Participants came from the same pool as that in Substudy 1. Data were collected from a total of 200 participants who did not participate in Substudy 1, with 100 participants randomly assigned to each of the two between-groups conditions. Across groups, participants did not differ significantly in age, class rank, or the number of jobs previously held.

Measures

The NEO Five-Factor Inventory (NEO-FFI; Costa & McCrae, 1989) was the personality measure used in this study. The NEO-FFI is a shortened form of the NEO-PI-R (Costa & McCrae, 1992), a measure based on the five-factor model of personality. The five factors of personality that the NEO-FFI measures are Neuroticism, Extraversion, Openness to Experience, Agreeableness, and Conscientiousness. The NEO-FFI consists of the 12 items having the highest positive or negative loading on each of the corresponding five factors of the NEO-PI-R. Responses to each item are recorded on a 5-point scale ranging from 0 (strongly disagree) to 4 (strongly agree). Previous estimates of internal consistency (coefficient alphas) for each of the 12-item scales were .86 (Neuroticism), .77 (Extraversion), .73 (Openness to Experience), .68 (Agreeableness), and .81 (Conscientiousness; Costa & McCrae, 1992). Costa and McCrae (1992) also reported that correlations between NEO-FFI scales and NEO-PI factors ranged from .75 to .89.

The second personality inventory used was an altered form of the NEO-FFI, on which a reference to work was appended to each statement, usually at the beginning or the end of the statement. For example, the item "I try to be courteous to everyone I meet" was modified to read "I try to be courteous to everyone I meet at work." Another example item was "I work hard to accomplish my work-related goals" instead of "I work hard to accomplish my goals." The Openness to Experience scale was not used in this study because work appendages did not make sense on many of the items.

Procedure

Substudy 1. Participants were randomly assigned to one of four groups and were told they would be completing a personality inventory. Written instructions were attached to each personality inventory. One group received an unaltered personality inventory (i.e., NEO-FFI) and was instructed to answer the questions as directed on the original version of the test (general instructions condition). Another group received the unaltered personality inventory, but their instructions indicated that they were to complete the inventory as if they were applying for a customer service representative job in a department store, a job they really wanted (applicant instructions condition). The remaining two groups received the altered personality inventory, one with general instructions and the other with applicant instructions.

Substudy 2. Participants were randomly assigned to one of
two groups (general instructions group or applicant instructions group). Participants in both groups completed both forms (counterbalanced within groups) of the personality test. Otherwise, all procedures were the same as those in Substudy 1.

**Results**

Participants' scores were calculated for each of the four NEO-FFI scales used. Table 1 reports the means and the standard deviations for each of these scales across the four conditions for both substudies. We then performed analyses of variance (ANOVA) on each of the NEO-FFI scales to test for mean score differences across the four experimental conditions in each study (i.e., Substudy 1: between-groups analysis; Substudy 2: repeated measures analysis with between-groups analysis on the instruction conditions). For Substudy 1, the analyses revealed no significant interaction between instruction type and item type, but there were main effects for both instruction type and item type on all scales except for the Extraversion scale. Extraversion scale scores were affected only by instruction type. F values, probabilities, and effect sizes for the main effects and the interactions are reported in Table 2.

For Substudy 2, the analyses showed a significant interaction between item type and instruction type for both the Neuroticism scale and the Conscientiousness scale (see Table 2). Consistent with Substudy 1, main effects for item type were found for the Neuroticism, Agreeableness, and Conscientiousness scales. Also consistent with Substudy 1, only a significant main effect for instruction type was found for the Extraversion scale. In Substudy 2, no significant main effect was found for instruction type for the Agreeableness scale. In general, the findings were very consistent across the two substudies, with the exception of the additional interactions that were found in Substudy 2.

The main effect findings confirmed the hypothesis that mean scale scores of individuals responding to work-specific personality items would be higher than those of individuals responding to noncontextual personality items. Also, as hypothesized, individuals in the applicant instructions condition had more positive scale scores than did those in the general instructions condition. The collapsed means and standard deviations for both substudies are presented in Table 3.

Consistent with the prediction made under the self-presentation theory of item responses, interactions between instruction type and item specificity were evidenced for two scales in Substudy 2. The nature of these interactions are depicted in Figures 1 and 2. As predicted, on both scales, individuals in the applicant instructions condition scored highest (i.e., in the most positive direction; lower scores are more positive on the Neuroticism scale) on the work-specific item scales, whereas individuals in the general instructions condition scored lowest on the noncontextual item scales.

The use of repeated measures designs in faking studies has been advocated by numerous researchers (e.g., Gordon & Gross, 1978; Lautenschlager, 1986). However, as demonstrated by Schwab (1971) and evidenced by Hough et al. (1990), the order of administration of the different response conditions often interacts with the effect of the response condition. This interactive effect was also found in the current study. An Item Type × Order of Administration interaction was found for the Conscientiousness scale, \( F(1, 196) = 9.47, p < .05 \); the Agreeableness scale, \( F(1, 196) = 8.11, p < .05 \); and the Neuroticism scale, \( F(1, 196) = 11.22, p < .05 \). The interaction for the Conscienc-
Table 2

<table>
<thead>
<tr>
<th>Scale</th>
<th>Instruction type</th>
<th>Item type</th>
<th>Instruction Type × Item Type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F</td>
<td>p</td>
<td>eta²</td>
</tr>
<tr>
<td>Neuroticism</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Substudy 1</td>
<td>14.82</td>
<td>.000</td>
<td>.13</td>
</tr>
<tr>
<td>Substudy 2</td>
<td>32.05</td>
<td>.000</td>
<td>.14</td>
</tr>
<tr>
<td>Extraversion</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Substudy 1</td>
<td>5.38</td>
<td>.023</td>
<td>.05</td>
</tr>
<tr>
<td>Substudy 2</td>
<td>12.69</td>
<td>.000</td>
<td>.06</td>
</tr>
<tr>
<td>Agreeableness</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Substudy 1</td>
<td>9.90</td>
<td>.002</td>
<td>.09</td>
</tr>
<tr>
<td>Substudy 2</td>
<td>1.62</td>
<td>.205</td>
<td></td>
</tr>
<tr>
<td>Conscientiousness</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Substudy 1</td>
<td>19.42</td>
<td>.000</td>
<td>.17</td>
</tr>
<tr>
<td>Substudy 2</td>
<td>20.16</td>
<td>.000</td>
<td>.09</td>
</tr>
</tbody>
</table>

The tests of the psychometric equivalence of the altered and unaltered forms of the measure followed the approach outlined by Jöreskog and others (Drasgow & Kanfer, 1985; Jöreskog, 1971a, 1971b). We conducted all of the analyses using LISREL 8 (Jöreskog & Sörbom, 1993). We used three indicators of fit to assess the models tested, including chi-square (and chi-square difference tests to examine loss of fit in nested models), root mean square error of approximation (RMSEA; Browne & Cudeck, 1993), and the comparative fit index (CFI; Bentler, 1990). An RMSEA of .05 has been suggested as an indicator of close fit, whereas .08 suggests a reasonable fit of the model to the data (Browne & Cudeck, 1993). CFI values greater than .90 also suggest a reasonable fit (Bentler & Bonett, 1980).

In all the LISREL analyses, the data from both substudies were included; thus, each of the four groups (i.e., noncontextual items, general instructions; noncontextual...
items, applicant instructions; work-specific items, general instructions; and work-specific items, applicant instructions) had a sample size of 125. Because of the relatively small sample size and the large number of items per common factor (i.e., 12 per scale, or 48 total), we adopted an approach used by Schmit and Ryan (1993) in the analysis of the factor structure of the NEO-FFI (cf. Drasgow & Kanfer, 1985). Accordingly, the 12 NEO-FFI items that comprise the four subscales used in the current study were randomly divided into three subsets, leaving a total of 12 item sets. Unit-weighted scores from these item sets were summed to create 12 composites. A $12 \times 12$ covariance matrix for each group was calculated and used in the analyses.

**Figure 1.** Item Type $\times$ Instruction Type interaction for the Conscientiousness scale. Scores on the Conscientiousness scale range from 0 to 48, with higher scores indicating more conscientiousness.

**Figure 2.** Item Type $\times$ Instruction Type interaction for the Neuroticism Scale. Scores on the Neuroticism scale range from 0 to 48, with higher scores indicating more neuroticism.
We conducted the first analyses to examine the one-versus four-factor structure issue for the altered personality items (i.e., items with "at work" tags). The data from the group in the general instructions condition with work-specific items were submitted to both a four-factor and a one-factor confirmatory factor analysis. In addition, the data from the group in the applicant instructions condition with work-specific items were submitted to the same analyses. The results are presented in Table 4. In both cases, it is clear that the four-factor model provided a better fit to the data than did the one-factor model.

The second set of analyses was a series of multiple group tests of psychometric equivalence between the noncontextual and work-specific item sets. In all analyses, four subtests were conducted with varying degrees of model restraint. The first test compared the form of the two models; that is, it was a test of whether the model for the two measures had the same number of latent variables with the same indicators and the same specification of fixed and free parameters. The second test constrained the latent-factor correlations to be invariant across measurement devices, and the third analysis further constrained the item sets to have equal loadings across the measures on the associated latent factors. The final analysis specified error variances to be equal across measures.

The first subset of analyses compared the equivalence of the altered items and the general items under the general instructions condition. The second subset of analyses compared the different types of items in the job-applicant instructions condition. The results of the two sets of four

Table 4
Comparison of One- and Four-Factor Models of Work-Specific Item Sets in General and Job-Applicant Instructions Conditions

<table>
<thead>
<tr>
<th>Model</th>
<th>$\chi^2$</th>
<th>$df$</th>
<th>$p$</th>
<th>RMSEA</th>
<th>CFI</th>
<th>$\chi_{\text{difference}}^2$</th>
<th>$df$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Four factor</td>
<td>98.68</td>
<td>48</td>
<td>&lt;.05</td>
<td>.09</td>
<td>.91</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>One factor</td>
<td>288.07</td>
<td>54</td>
<td>&lt;.05</td>
<td>.19</td>
<td>.58</td>
<td>189.39</td>
<td>6</td>
<td>&lt;.05</td>
</tr>
</tbody>
</table>

General instructions

| Four factor    | 107.76  | 48   | <.05 | .10   | .92   |                             |      |      |
| One factor     | 288.07  | 54   | <.05 | .17   | .75   | 138.54                      | 6    | <.05 |

Job-applicant instructions

Note. For the chi-square values, $N = 125$. RMSEA = root mean square error of approximation; CFI = comparative fit index.
models tested are presented in Table 5. In both cases, psychometric equivalence was strongly supported. Under both conditions, invariant structure, invariant latent-factor correlations, and invariant factor loadings were evidenced across measures. Only error variances were not invariant across measures. An examination of the error variances suggested that in both the general instructions condition and the job-applicant instructions condition, the error variances were significantly larger for the noncontextual items (.31 for general instructions condition and .28 for job-applicant instructions condition) than for the work-specific items (.28 for general instructions condition and .26 for job-applicant instructions condition), \( F_{max}(124) = 1.12, p < .05 \), and \( F_{max}(124) = 1.08, p < .05 \), respectively.

**Discussion**

The tests of the four-factor structure and the psychometric equivalence provided evidence that the altering of items to include “at work” tags might not have adversely affected the structure or other psychometric properties of the scales. In light of these findings, the scale comparisons made in the two substudies make sense from both a practical and a conceptual level.

All main effects of the experimental manipulations were in the expected direction; participants in both the job-applicant instructions and work-specific item conditions had more positive scale scores than did participants in the general instructions and noncontextual item conditions, respectively. No interactions between instruction type and item type were found in Substudy 1, but interactions for two of the four scales were found in Substudy 2.

The findings suggest that differences in personality-inventory scale scores are affected by the frame of reference that the respondent considers when completing the measures. As hypothesized, both the specificity of the context in which the inventory is completed and the specificity of the items appear to affect the scale scores. However, the combination of the two variables also appears to affect at least some scale scores (e.g., Conscientiousness and Neuroticism). This finding is consistent with the prediction based on the theory of self-presentation in item responses. The only scale on which no score differences were related to item type was the Extraversion scale. This may suggest that this particular scale generalizes better across situations than do the other scales used in the study. Alternatively, this scale may represent a set of behaviors less constrained by the workplace or less subject to social desirability than the other scales used in the study; however, given the fictitious job used in the current study, customer service representative, these alternatives seem unlikely.

Somewhat larger effect sizes for the instruction type variable than for the test type variable were found in most comparisons, and their pervasiveness across all scales in both substudies suggests that social desirability may be at least a partial explanation for score differences. Furthermore, the interactions between instruction type and item type may be suggestive of social desirability effects rather than self-presentation effects. As noted earlier, socially desirable responding may be a detriment to validity, whereas self-presentation effects serve to increase validity. Thus, we designed a second study with the objective of trying to identify which may be the better explanation for score differences.

**Study 2**

The primary objective of the second study—a between-groups criterion-related validity study—was to compare the validities of altered and unaltered tests in the two instruction conditions. That is, does the situation, the item context, or both affect the criterion-related validity of per-
sonality tests? A secondary objective was to explore the role played by social desirability and self-presentation effects on the criterion-related validities. As noted earlier, both socially desirable responding and frame-of-reference confusion should lead to greater error variance in the prediction of a criterion by a personality test. Thus, if situation-induced social desirability (through both instructions and items) plays a role in response differences, validity should be significantly smaller (or nonsignificant) in the applicant instructions–work-specific item condition than in the other three conditions. If item social desirability alone affects responses (such that variability is reduced through a negative skew of the distribution), then the validity in the general instructions–work-specific item condition should be lower than that in the general instructions–noncontextual item condition. However, if the self-presentation argument is correct, then validity should be higher in the applicant instructions–work-specific item condition than in the applicant instructions–noncontextual item condition, because all respondents will be using the same or more similar frames of reference, thereby reducing error variance. Furthermore, the theory of conditional dispositions suggests that the validity of the specific items should be higher than that of the noncontextual items in the general instructions condition, because individuals are posited to be more predictable within specific situations (e.g., at work) than across all situations. Finally, the conditional disposition and self-presentation theories considered in tandem would rank order validity predictions, from highest to lowest as (a) applicant instructions, work-specific items, (b) general instructions, work-specific items, (c) general instructions, noncontextual items, and (d) applicant instructions, noncontextual items.

The Conscientiousness factor of the Big Five personality factors has been shown to be a useful predictor in most jobs (Barrick & Mount, 1991; Hough et al., 1990; Tett et al., 1991). In addition, Hough et al. specified both predictors and criteria in a meta-analysis of personality variables and found that both dependability and achievement measures were useful predictors of educational criteria (uncorrected criterion-related validity coefficients were .15 and .30, respectively). These two variables are included as facets of the Conscientiousness scale of the NEO-PI-R (Costa & McCrae, 1992), a measure based on the five-factor model of personality. In the current study, the Conscientiousness scale and an altered school-specific version of it were used to predict students’ grade point averages (GPAs). Previous research has shown the Conscientiousness scale to be a useful predictor of college course grades, with a criterion-related validity of .25 (Dollinger & Orf, 1991).

Method

Participants

Students of an introductory psychology course participated for course credit. Data were collected from a total of 400 participants, with 100 participants randomly assigned to each of the four study conditions; data from 6 participants were incomplete and were dropped before data were analyzed. Across groups, participants did not differ significantly in age, class rank, or the number of jobs previously held.

Measures

NEO-PI-R Conscientiousness scale. The Conscientiousness scale of the NEO-PI-R (Costa & McCrae, 1992) was the personality measure used in the current study. The Conscientiousness scale consists of 48 items measuring six facets, including Competence, Order, Dutifulness, Achievement Striving, Self-Discipline, and Deliberation. Responses to each item are coded on a 5-point scale ranging from 0 (strongly disagree) to 4 (strongly agree). Previous estimates of internal consistency (coefficient alphas) for each of the facet subscales were .67 (Competence), .66 (Order), .62 (Dutifulness), .67 (Achievement Striving), .75 (Self-Discipline), and .71 (Deliberation), while the entire Conscientiousness scale had an alpha of .90 (Costa & McCrae, 1992).

The second personality inventory used was an altered form of the Conscientiousness scale, on which a reference to school was appended to each statement, usually at the beginning or the end of the statement. For example, the item “I strive for excellence in everything I do” was modified to read “I strive for excellence in everything I do at school.”

Criterion. The criterion measure used was college cumulative GPA. Permission was obtained from the students to gain access to the registrar’s records containing their college GPAs. Six individuals refused to give this permission, and their predictor data were dropped from the analysis.

Procedure

Participants were randomly assigned to one of four groups and were told they would be completing a personality inventory. Written instructions were attached to each personality inventory. One group received an unaltered personality inventory and were instructed to answer the questions as directed on the original version of the test (general instructions condition). Another group received the unaltered personality inventory, but their instructions asked them to imagine that they had just arrived at the admissions office of a university they really wanted to attend. They were then told that admission decisions would be based in part on their performance on the personality test they were about to complete. Finally, they were told that those students meeting the qualification standards on the test would receive prize money ($10) in lieu of the admission. The remaining two groups received the altered personality inventory, one with general instructions and the other with school-applicant instructions, which also included the prize money (i.e., $10) incentive clause. Finally, following the completion of the test, participants were asked for their consent to have the registrar release their GPAs to the researchers.

The prize money incentive provided in the two school-applicant instructions conditions was used to induce motivation in the students similar to motivational forces that might be found in actual applicants. After all testing was completed, participants in all four conditions were debriefed with a letter stating...
the purpose and the hypotheses of the study. The need for motivating some participants with a reward incentive was explained. Participants were also informed that a random drawing would take place to distribute fifteen $10 prizes, the total of the pool of potential money promised earlier to some participants as motivating incentives.

**Results**

We calculated criterion-related validity coefficients for both the broad predictor scale (the Conscientiousness scale) and the more narrow facet subscales using GPA as the criterion. The validity coefficients were calculated for each of the four experimental groups. The results are presented in Table 6. Also included in Table 6 are the alpha reliability coefficients for each scale, calculated within group. With the exception of the reliabilities for the Dutifulness subscale, which were low and variable across conditions, the scales shared very similar reliability estimates across conditions.

The rank order of the validity coefficients for the Conscientiousness scale across the four conditions was consistent with the predictions made by the conditional disposition and self-presentation theories: (a) applicant instructions, school-specific items \((r = .46, p < .01)\), (b) general instructions, school-specific items \((r = .41, p < .01)\), (c) general instructions, noncontextual items \((r = .25, p < .01)\), and (d) applicant instructions, noncontextual items \((r = -.02, ns)\). In support of self-presentation theory, the validity of the school-specific items \((r = .46)\) was significantly higher in the applicant instructions condition than the validity of the noncontextual items \((r = -.02), z = 3.30, p < .05\). In further support was the finding of significantly higher validity in the general instructions–school-specific item condition \((r = .41)\) than in the applicant instructions–noncontextual item condition \((r = -.02), z = 3.15, p < .05\). The conditional disposition hypothesis received more support than the item social desirability hypothesis, as the validity of the school-specific items \((r = .41)\) was greater than the validity of the noncontextual items \((r = .25)\) in the general instructions condition. Although the validity difference was in the direction that supported the condition disposition hypothesis, the difference was not significant, \(z = 1.25, ns\). Finally, the validities for school-specific items did not differ significantly between the applicant instructions condition \((r = .46)\) and the general instructions condition \((r = .41), z = 0.43, ns\); the validities of the noncontextual items were also not significantly higher in the general instructions condition \((r = .25)\) than in the applicant instructions condition \((r = -.02), z = 1.94, ns\).

In general, these findings at the broad level of measurement were also repeated at the more narrow facet level of measurement (see Table 6).

In an attempt to replicate the findings of Study 1, we performed a between-groups ANOVA to identify mean Conscientiousness scale score differences across conditions. As was the case in Substudy 1, significant main effects for instruction type, \(F(1, 390) = 23.94, p < .05\), and item type, \(F(1, 390) = 5.56, p < .05\), were found; however, the interaction was not significant, \(F(1, 390) = 0.68, ns\). Individuals in the applicant instructions conditions \((M = 129.86, SD = 21.95)\) scored significantly higher than individuals in the general instructions conditions \((M = 119.47, SD = 20.40)\). Individuals in the school-specific item conditions \((M = 127.18, SD = 22.39)\) scored significantly higher than those in the noncontextual item conditions \((M = 122.17, SD = 20.95)\).

A confirmatory factor analysis was also conducted in this study in a manner similar to that in Study 1. That is, the psychometric equivalence of the measures across the four independent groups was assessed. The developers of the Conscientiousness scale intended each of the six facet

---

### Table 6

*Criterion-Related Validities for Grade Point Average and Reliabilities of Conscientiousness Scale and Subscales Across Experimental Conditions*

<table>
<thead>
<tr>
<th>Scale</th>
<th>General, noncontextual</th>
<th>Applicant, noncontextual</th>
<th>General, school-specific</th>
<th>Applicant, school-specific</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(r)</td>
<td>(\alpha)</td>
<td>(r)</td>
<td>(\alpha)</td>
</tr>
<tr>
<td>Conscientiousness</td>
<td>.25*</td>
<td>.91</td>
<td>-.02</td>
<td>.90</td>
</tr>
<tr>
<td>Competence</td>
<td>.31**</td>
<td>.71</td>
<td>-.02</td>
<td>.68</td>
</tr>
<tr>
<td>Order</td>
<td>.02</td>
<td>.64</td>
<td>-.08</td>
<td>.55</td>
</tr>
<tr>
<td>Dutifulness</td>
<td>.19</td>
<td>.34</td>
<td>-.08</td>
<td>.42</td>
</tr>
<tr>
<td>Achievement Striving</td>
<td>.25*</td>
<td>.81</td>
<td>-.10</td>
<td>.68</td>
</tr>
<tr>
<td>Self-Discipline</td>
<td>.17</td>
<td>.79</td>
<td>.06</td>
<td>.79</td>
</tr>
<tr>
<td>Deliberation</td>
<td>.23*</td>
<td>.77</td>
<td>.10</td>
<td>.72</td>
</tr>
</tbody>
</table>

\*\(n = 99\).  \*\(n = 98\).

\*\(p < .05\)  **\(p < .01\).
subscales of the broader measure to have substantial loadings on the underlying trait of conscientiousness (Costa, McCrae, & Dye, 1991). Accordingly, each of the six facet subscales was hypothesized to have significant independent loadings on a single latent variable. The test of similar factor structure across groups showed a close fit between the data and the model, \( \chi^2(36, N = 394) = 41.71, p = .24, \text{RMSEA} = .02, \text{CFI} = 1.00 \). With the additional constraint of equal factor loadings across groups, the model still fit well, \( \chi^2(54, N = 394) = 67.83, p = .10, \text{RMSEA} = .03, \text{CFI} = .99, \) and no significant loss of fit was indicated, \( \chi^2(18, N = 394) = 26.12, p > .05 \). In addition, as hypothesized, all six facet subscales had significant loadings on the latent variable. The third model tested constrained error variances to be equal across groups. As in Study 1, this model did not fit the data as well as the previous model, \( \chi^2(72, N = 394) = 99.22, p = .02, \text{RMSEA} = .07, \text{CFI} = .98, \) and resulted in a significant loss of fit, \( \chi^2(18, N = 394) = 31.39, p < .05 \). Also parallel to Study 1, the error variances were found to be significantly larger for the noncontextual items (.43) than for the school-specific items (.36), \( F_{\text{max}}(199) = 1.19, p < .05 \).

Discussion

In Study 2, the findings offer more support for the conditional disposition and self-presentation hypotheses than for the social desirability hypothesis. Validity was highest in the condition that used context-specific items in the applicant situation; conversely, validity was essentially zero when general context items were used in the applicant condition. As Johnson (1981) hypothesized, making self-presentation easier by using work-specific items appeared to increase validity. In other words, altering the items gave all respondents a common frame of reference, which appears to have reduced error variance and increased validity.

In Study 2, the significant mean differences in scores between conditions found in Study 1 were replicated; participants in both the applicant instructions and school-specific item conditions had more positive conscientiousness scale scores than did participants in the general instructions and noncontextual item conditions, respectively. The finding of psychometric equivalence of the measures across conditions was also similar to the findings of Study 1. Thus, replication of results across studies was evidenced.

Earlier we suggested that the consequence of job applicants using different frames of reference was an increase in the error of prediction. This proposition was supported but not fully explained in the current study. A portion of the error is likely to be nonrandom, resulting in a moderator variable. One moderator of personality-test validity already identified, the self-monitoring variable (Snyder & Ickes, 1985), may help explain why applicants use different frames of reference. High self-monitors tend to rely on the immediate situation to guide their behavior more than do low self-monitors. Consequently, high self-monitors given a noncontext-specific personality test in a personnel selection situation are probably more likely than low self-monitors to respond to personality-test items using work-related experiences as a guide. However, if personality-test items referred to a work-related context, the difference between high and low self-monitors would likely be attenuated. Future research is needed to determine whether self-monitoring is the moderator that may explain the results found in the current study or whether some other moderator may be responsible (see Chaplin, 1991); for example, scalability is a potential alternative moderator (Lanning, 1988).

General Discussion

In both Studies 1 and 2, the mean differences in personality-test scores were consistent with the conditional disposition and self-presentation hypotheses, although social desirability effects could not be ruled out as an alternative explanation by these findings alone. However, the criterion-validity differences across conditions in Study 2 clearly supported the self-presentation theory hypothesis and, to a lesser extent, the conditional disposition theory hypothesis over the alternative social desirability hypothesis. The distinction here is between positive self-presentation alone and positive and accurate self-presentation. If individuals present themselves in a positive light, but they do so inaccurately, that is social desirability. This form of measurement error should result in lower validity or possibly no change in validity if few individuals do it (e.g., Hough et al., 1990). If individuals present themselves positively and accurately because they have a frame of reference, then validity should increase. To expect job applicants not to engage in some positive impression management is unrealistic (Leary & Kowalski, 1990; Schlenker & Weigold, 1992), regardless of the method of measurement. The key is to find ways to help them present themselves more accurately given the job context; the current study suggests a way to do that.

Although replication in a field setting is still needed, the practical implications for personnel selection are clear. Face-valid items do appear to have the potential for realizing higher criterion-related validity (and consequently, increased utility) when used for personnel selection. Rynes (1993) recently made a plea to researchers to test the idea that increases in the face validity of a test may be related to increases in criterion-related validity of the test. Others have noted both that face validity has been a "well kept secret in the empirical tradition of test development" and
that "the best items (from an empirical perspective) tend to be the ones with good face validity" (Hogan, Carpenter, Briggs, & Hansson, 1985, p. 30). This seems to be true at least for the Conscientiousness scales used in the current study.

Future research is needed to test whether changes in face validity (i.e., item context changes) increase the criterion-related validity of other measures of personality. For example, mean differences across item type conditions were not found for the Extraversion scale in both Substudies 1 and 2. This finding leaves open the possibility that validity differences across conditions comparable to those found for the Conscientiousness scale in Study 2 may not be found for a measure of extraversion. The conscientiousness trait has been found to generalize across most jobs (Barrick & Mount, 1991), and this likely includes the "job" of being a college student. Thus, we probably have optimized our chance for generalizability of the Study 2 findings by using only this trait, and other trait measures may not show the same improved validities associated with simple frame-of-reference changes. Indeed, moderator research in personality psychology has found the domain of conscientiousness the most fruitful for demonstrating moderator effects; other domains have been less promising (Chaplin, 1991).

More research also needs to be done with variables at a bandwidth more narrow than the Big Five (Rothstein, Jackson, & Tett, 1994). In the current study, the Order facet of Conscientiousness was not a useful predictor in any condition. In addition, the findings for the Dutifulness facet were less consistent with the overall findings for the Conscientiousness scale than was the case for the remaining four facets. Lower reliability for these scales, however, explains much of their lower and inconsistent criterion-related validity in the current study. Future research with other Big Five constructs and the associated facets may produce different results than those found for Conscientiousness.

It appears that the use of general personality items in a personnel selection context does result in significantly more error on the predictor side of the predictor–criterion equation, resulting in lower validities than is the case for context-specific items. This was evidenced by the lower error variances for the work-specific item sets than for the noncontextual item sets in both studies. This additional predictor error is also likely to affect the factor pattern of personality measures used in a personnel selection context. Indeed, Schmit and Ryan (1993) found that a simple factor pattern for a set of personality measures completed by a volunteer sample was substantially different from a complex factor pattern found in an applicant sample. Furthermore, the complex factor pattern of the applicant sample included a primary factor that the authors labeled as an ideal employee factor. This factor included most of the items on the personality form that were most directly related to work. The self-presentation item response theory would suggest that these were the items on which self-presentation was easiest to do; these items would also be most consistent with the item development advice of Johnson (1981). In Study 1, we found evidence for psychometric equivalence of the two test forms within applicant and nonapplicant conditions. A test of applicant versus nonapplicant differences, as was presented by Schmit and Ryan (1993), was not possible because of the nonindependence of the group combinations that would make this test possible. However, a post hoc examination of differences in average factor intercorrelations in Study 1 revealed that higher factor intercorrelations were present in the job-applicant instructions condition than in the general instructions condition for both noncontextual and work-specific items (results are available from Mark J. Schmit).

The higher factor intercorrelations in the applicant groups are consistent with Paulhus, Bruce, and Trapnell's (1995) findings that in the job-applicant instructions condition, even when told to fake being the best candidate for the job, some individuals respond in a purely self-descriptive manner. Paulhus et al. showed that this led to inflated factor correlations in a condition in which restriction of range might be expected. Inflated correlations in an applicant population may also come from applicants using different frames of reference (i.e., general vs. work specific). However, if this were the case, factor correlations would be expected to drop when frame of reference is held constant with work-specific items; this was not found to be true in the current study. It could be that when work-specific items are used, factor intercorrelations are increased because individuals have clear conceptions about the social norms that call for more consistency of trait behaviors within certain situations, such as work-related situations, than across all situations (cf. Mischel, 1973; Moskowitz, 1994). Future research is needed to explore this possibility.

Clearly, the greatest limitation of the current studies is the potential for limited external validity because of the fact that they were simulated selection-context lab studies. The replication of findings within the set of studies is a positive indicator for the possibility of external validity, but field replication is still needed. For example, in Study 1 we found psychometric equivalence for both forms of the test items in the job-applicant instructions condition. Although the factor correlations in this study were inflated in the applicant instructions condition, as in that of Schmit and Ryan (1993), the complex factor loading found by Schmit and Ryan in a real applicant population was not evidenced. In Study 2, a monetary incentive was used in the applicant instructions condition, but this may have fallen short of having the motivating potential of a
real position opening. Finally, Study 2 involved an academic predictor and criterion that may not have the same generalizability to work predictor–criterion relationships. Field replication should address these issues. This field replication should also explore the possibility that even more narrowly defined situational contexts may be required for some jobs in the writing of personality items to potentially increase validity. For example, the item “I am courteous to everyone I meet at work” might be made more situationally relevant by changing it to “I am courteous to every customer I meet at work.” An individual in a customer service position may be very courteous to customers but very rude to coworkers. If the behavior of most value to the organization is related to customer service and if the criterion reflects this, then the second item may be more predictive. It may be that items from scales, other than possibly the Conscientiousness scale, require different levels of context specificity to make them more valid than the general items when used for personnel selection. Finally, a logical alternative to altering every test item to be context specific would be to simply instruct the respondents to answer all questions with respect to the work context; although, as shown by Paulhus et al. (1995), some respondents may ignore instructions and answer the items in the general context to which they are referenced.

Given the recent increased interest in the use of personality measures in personnel selection, it is important that researchers continue to look for ways to improve the predictive efficiency of these types of tests. This set of studies represents a first step in that direction.

1This suggestion was made by an anonymous reviewer.

References


Shoda, Y., Mischel, W., & Wright, J. C. (1989). Intuitive inter-

Received December 20, 1994
Revision received April 20, 1995
Accepted April 20, 1995


