The Viability of Retesting for Managing Faking Tendencies

Jill E. Ellingson
The Ohio State University

Eric D. Heggestad
Colorado State University

February 2004

An earlier version of this paper was presented in E. D. Heggestad (Chair), Practical Considerations for Implementing Personality Testing in Organizations. Symposium conducted at the annual meeting of the Society for Industrial and Organizational Psychology, Orlando, FL. April 2003

Author Note

Jill E. Ellingson, Department of Management and Human Resources; Eric. D. Heggestad, Department of Psychology.

Correspondence concerning this article should be addressed to Jill E. Ellingson, Department of Management and Human Resources, 734 Fisher Hall, 2100 Neil Avenue, Columbus, OH, 43210. Electronic mail may be sent to ellingson@cob.osu.edu.
Abstract

Personality measures often include impression management scales to identify individuals who are faking. This study investigated whether retesting those individuals represents a viable corrective technique for managing faking tendencies. The evaluation of retesting was paired with an investigation of warnings as a potentially effective prevention technique. Further, because both techniques may impact applicant attitudes about the assessment process, test-taker attitudes associated with retesting and warnings were investigated. The results provided supportive evidence for the impact of retesting on faking behavior, but also suggested that asking individuals to retake a personality measure can result in negative attitudes about the assessment. The warning results reaffirmed previous research suggesting that warnings, without the power of consequences, do little to impact faking behavior.
The Viability of Retesting for Managing Faking Tendencies

There is growing empirical evidence that personality measures are valuable predictors of job performance (e.g., Barrick & Mount, 1991; Hough, 1992; McHenry, Hough, Toquam, Hanson, & Ashworth, 1990; Mount & Barrick, 1995; Salgado, 1997) and are therefore useful when making hiring decisions. However, discussions regarding the viability of personality measurement within organizational contexts often lead to concerns about respondent faking.¹ As self-report measures, personality inventories have the potential to be influenced by socially desirable responding. Most organizations view this potential as unavoidable, a perspective that significantly reduces the instrumentality of these measures as decision-making tools. Given these fears about faking, any recommendation to include personality assessment within selection systems must be paired with suggestions regarding how to manage socially desirable responding.

Hiring and promotion decisions generally occur at the level of the individual. For example, a common means of differentiating between applicants involves ranking them on the basis of their observed personality scale scores. A top-down selection rule or cut-off point is then used to select those applicants who display more desired characteristics. Individuals who distort their responses to appear more desirable will increase their ranking (except when faking is universal and all applicants are equally adept at faking), thereby displacing those individuals who responded honestly. Faking in this context can have significant effects on hiring decisions (Christiansen, Goffin, Johnston, & Rothstein, 1994; Mueller-Hanson, Heggestad, & Thornton, 2003; Rosse, Stecher, Miller, & Levin, 1998; Zickar, Rosse, Levin, & Hulin, 1999).

A variety of methods have been proposed to reduce faking on self-report assessments. Specifically, the construction of forced-choice instruments, empirically-keyed instruments containing more subtle items, and the application of corrections to personality trait scale scores
have all been explored as methods for managing faking on personality measures (e.g., Christiansen et al., 1994; Ellingson, Sackett, & Hough, 1999; Hough, 1998; Jackson, Wroblewski, & Ashton, 2000; Paulhus, 1981). However, research on the effectiveness of these approaches has been inconclusive at best. In fact, recent discussions concerning the use of corrections go so far as to argue that any application of this approach is questionable given current empirical evidence (Goffin & Christiansen, 2003). Thus, organizations continue to lack effective strategies for dealing with individuals who distort their responses.

Retesting

This study was designed to evaluate the use of retesting as an alternative for managing faking on personality measures. Many personality measures designed for organizational use (e.g., California Psychological Inventory, Sixteen Personality Factor Questionnaire (16PF), Global Personality Inventory, Personality Research Form) include an impression management scale as a means of tracking those individuals who may have faked their responses (e.g., Conn & Rieke, 1994; Gough & Bradley, 1996; Jackson, 1999; Schmit, Kihm, & Robie, 2000). Test manuals commonly recommend that users rely on these scales as an index of the extent to which individuals may be distorting their responses. When an individual is “flagged” by the scale, the manuals often suggest that this individual’s scores should be cautiously interpreted. Such an arms-length response on the part of the assessment firms is understandable given that establishing decision-making policy for organizations would be outside of their purview. Yet, such direction constitutes little advice regarding what to do with red-flagged individuals, leaving organizations to contemplate a reasonable course of action on their own.

Having flagged an otherwise qualified candidate, an organization has several possible courses of action. First, the organization could choose to eliminate the individual(s) from further
consideration. This option is clearly a difficult choice in those contexts where the organization faces a small applicant pool, has invested valuable resources in recruiting, and/or the red-flagged individual(s) has performed well on the other aspects of the decision-making process. Second, the organization could choose to disregard the personality scale scores for those individuals who are red-flagged. Yet, this choice would impact the extent to which individuals could be compared in a standardized manner. Third, the organization could choose to disregard the results from the impression management scale itself. However, the process of questioning the veracity of one scale from a personality measure necessarily leads to questioning the veracity of the measure as a whole. Further, research by Mueller-Hanson et al. (2003) indicates that keeping applicants with high scores on an impression management scale in the applicant pool can adversely impact the utility of the personality instrument. Fourth, the organization could request that these individuals retake the test and encourage them to respond more honestly the second time. This retesting approach has been advocated in test manuals. For example, when interpreting high scores on its impression management scale, the 16PF manual suggests that, “Depending on the reasons for testing and the criticality of accurate test data, the professional might consider retesting, especially if deliberate distortion is expected” (Russell & Karol, 1999).

Of these options, retesting may be the most viable course of action for responding to individuals that have been red-flagged when personality measures are used for decision-making purposes. Specifically, indicating to individuals that they have been identified by an impression management scale as responding in a less than honest manner and allowing them to retake the test with instructions to respond more honestly the second time should reduce faking. Assuming that response patterns are altered accordingly, decisions concerning the individuals could be made using a more honest set of scores.
In addition to serving as a more palatable option, retesting red-flagged individuals offers other potential advantages for organizations. Retesting can emphasize to respondents the importance of answering the questions in an honest manner. Since research suggests that the base rate of red-flagged individuals is low (Ellingson, Smith, & Sackett, 2001; Gough & Bradley, 1996), retesting should be relatively feasible and affordable. In addition, providing individuals with an opportunity to retake the test would be in keeping with the Uniform Guidelines on Employee Selection Procedures (1978), which emphasize the importance of retesting and reconsideration, when reasonable, in an attempt to provide individuals with a fair and impartial assessment experience.

However, the advantages of retesting red-flagged individuals are meaningful only to the extent that we have a clear understanding of the implications associated with informing applicants that their responses appeared suspicious and asking them to complete the measure for a second time. Unlike cognitive ability testing where retesting research occurs regularly under the guise of investigating practice effects (e.g., Hausknecht, Trevor, & Farr, 2002; Kulik, Kulik, & Bangert, 1984; Sackett, Burris, & Ryan, 1989), little research has gone beyond simple reliability measurement to document the implications of having individuals retake a personality measure. One exception is a study by Kelley, Jacobs, and Farr (1994) that investigated the impact of multiple administrations of the Minnesota Multiphasic Personality Inventory (MMPI). Their research documented a trend toward normality in MMPI scale profiles over the course of multiple assessments. While the mechanism underlying this trend was unclear, the scale score differences observed provide support for the notion that retesting is likely to produce a meaningful change in scale scores.
Warning Statements

When managing faking tendencies, it is clear that an organization may choose to prevent it from occurring initially, or an organization may choose to detect those who have distorted their responses and correct those responses accordingly. In actuality, an organization that leverages both approaches has the best opportunity for effectively managing faking. The two approaches should work in conjunction wherein the front-end technique decreases the initial extent of faking, while the back-end technique gives the organization a means of dealing with those individuals who continue to distort their responses. Since retesting is an example of a detection-and-correction technique, a thorough evaluation of its viability requires that one consider how retesting works when used in conjunction with a preventative technique. Warnings represent a common prevention strategy for use when managing faking tendencies. With this technique, respondents are reminded of the importance of responding honestly on the personality measure and warned of the potential for detection should one fail to do so. In contrast to retesting, the effectiveness of warnings for reducing faking tendencies has been studied quite regularly. Dwight and Donovan (2003) provided meta-analytic results suggesting that the presence of warnings does decrease the extent of applicant faking. Specifically, personality measure scores obtained when respondents are subject to warnings are 0.23 standard deviation units lower on the average than when respondents are not subject to warnings. Given these supportive results, we sought to further our evaluation of retesting by incorporating an evaluation of warnings as well.

Respondent Reactions

The use of retesting and warnings when administering personality measures represents a procedural change in the assessment process. This change may impact how applicants react to the assessment experience. Personality measures tend to be associated with more negative
applicant reactions relative to other selection techniques (e.g., Rosse, Miller, & Stecher, 1994; Steiner & Gilliland, 1996). Asking individuals to retake a personality measure because they were red-flagged and/or informing individuals that their performance on a personality test is being monitored through a faking scale may result in applicant perceptions about the assessment experience that are even more negative. Alternatively, the provision of a chance to retake the test may produce more positive applicant reactions to the extent that the opportunity for reconsideration is viewed by the applicant as a more fair response on the part of the organization (e.g., Ployhart & Ryan, 1998; Truxillo, Bauer, Campion, & Paronto, 2002). Since personnel selection research suggests that applicant attitudes about the assessment process have the potential to impact valued organizational outcomes such as job satisfaction, willingness to recommend the organization to others, and intentions to accept a job offer (e.g., Gilliland, 1994; Macan, Avedon, Paese, & Smith, 1994; Rynes & Connerly, 1993), any evaluation of the viability of retesting and warnings should consider not only how each technique impacts faking tendencies, but also whether each technique impacts important attitudinal outcomes stemming from the assessment experience. Thus, the impact of retesting and warnings on the extent to which individuals were satisfied with the assessment experience, the extent to which individuals felt anxious about the assessment, the extent to which individuals viewed the test as a fair measure of their personality, and the extent to which individuals were motivated to complete the measure was investigated in an exploratory manner.

Method

Sample

The sample consisted of 235 undergraduate students enrolled in an introductory psychology course at a large university in the western United States. The sample was 68%
female and 88% White. The mean age of the sample was 19 years (SD = 1.5). Forty-one percent of the sample was employed at the time of the study either in full-time or part-time work, and thirty percent of the sample was familiar with personality inventories through being asked to complete a personality measure when applying for a job. For those participants, 3% received feedback from the organization about how they had scored on the measure they completed.

Measures

**Personality.** The personality measure used in this study was the Key Point Job Fit Assessment. The Job Fit is a 60-item personality inventory that is commercially available, and thus represents a measure that organizations use for applied purposes. Each of the Job Fit items is presented within a work frame-of-reference to enhance the perceived relevance of the assessment. Participant responses are measured on a 5-point Likert-type scale.

The Job Fit is designed to measure five personality traits specifically identified through previous research as important for job performance (Lahti, Mueller-Hanson, Heggestad, & Hastey, 2001). The Achievement Striving scale measures the tendency to work hard, a willingness to delay personal gratification to meet work goals, and an intrinsic desire to improve one’s work-related skills. The Assertiveness scale measures preferences toward dealing with others directly, expressing opinions openly, and taking initiative in work situations. The Dependability scale measures the tendency to be punctual, reliable, and responsible at work. The Extroversion scale measures preferences for working with others and interacting with others. The Stress Tolerance scale measures how an individual handles stress at work. The Job Fit also measures socially desirable responding through the Candidness scale, which measures the extent to which an individual’s responses may have been altered by a desire to present a positive impression. Test developers indicate that organizations should “interpret with caution” (pp. 9)
any assessment results that are obtained from individuals who receive a high score (i.e., roughly 1.5 standard deviations above the mean) on the Candidness scale.

Recent research has provided evidence for the reliability and construct validity of the Job Fit (Hastey, Heggestad, Mueller-Hanson, & Lahti, 2001). In the present study, alpha coefficients demonstrated acceptable internal consistency reliability across all scales for both the Time 1 and Time 2 administrations of the test: Achievement Striving scale, 0.76 and 0.76; Assertiveness scale, 0.75 and 0.83; Dependability scale, 0.68 and 0.72; Extroversion scale, 0.73 and 0.69; Stress Tolerance scale, 0.72 and 0.76; Candidness scale, 0.79 and 0.74.

Test Satisfaction. Satisfaction with the Job Fit was measured using a three-item scale adapted from Tonidandel, Quinones, and Adams (2002). An example item from this scale is, “This questionnaire was appealing to me.” The alpha coefficients for this scale at Time 1 and Time 2 were 0.82 and 0.84.

Perceived Test Fairness. Perceptions about the fairness of the assessment were measured using a five-item scale adapted from Quinones (1995). An example item from this scale is, “This questionnaire obtains accurate information about each person’s personality.” The alpha coefficients for this scale at Time 1 and Time 2 were 0.88 and 0.91.

Test Anxiety. Participant anxiety in response to the assessment was measured using an 11-item scale adapted from Tonidandel et al. (2002). An example item from this scale is, “This questionnaire made me feel a little tense.” The alpha coefficients for this scale at Time 1 and Time 2 were 0.85 and 0.84.

Test-taking Motivation. Participant motivation to complete the assessment was measured using a seven-item scale adapted from Arvey, Strickland, Drauden, and Martin (1990). An
example item from this scale is, “Doing well on this questionnaire was important to me.” The alpha coefficients for this scale at Time 1 and Time 2 were 0.76 and 0.84.

**Procedure**

A repeated-measures design was used to assess the change in Job Fit and attitude scale scores when individuals were asked to take the measure again for the purpose of obtaining more honest responses. Students were invited to participate in a two-session study for partial fulfillment of a course requirement. In the first session (Time 1), participants were asked to respond to the Job Fit as if it was being administered as part of an application process for a desirable position. These motivating instructions were used to create the demand characteristics typical of high-stakes decision-making settings. After completing the Job Fit, the participants were asked to complete the attitude scales as a measure of their reactions to the experience of completing the personality questionnaire. Following the first session, the participants were separated into two groups based on their Candidness scale score: those participants who scored within the normal range (i.e., no flag group) and those who scored above a preset criterion (i.e., flagged group).²

After 2-3 days, both groups returned for a second session (Time 2). Participants in the flagged group were informed that a social desirability scale detected that they were likely distorting their responses, and they were asked to complete the Job Fit again, more honestly. Participants in the no flag group were told that due to a computer failure, the data from their previous assessment was lost. They were asked to please complete the Job Fit again. The no flag group served as a control group; comparative changes in their scores would reflect the impact of test-retest reliability effects or simply the frustration of having to complete the assessment again. The same motivating instructions were used in the second administration of the Job Fit as were
used in the first administration. Finally, both groups were asked to complete the attitude scales again, followed by a series of demographic questions.

A warning manipulation was introduced as an overlay to this design. Half of the participants completed the two sessions under a warning condition, and the other half completed the two sessions with no warning. In the warning condition, participants were informed that there were items in the test that could be used to detect when an individual is responding in an overly positive manner. The participants were told of the importance of responding honestly, and that responses to these particular items will signal if they are responding less than honestly. In the no warning condition, participants were not informed about the presence of these items. The warning manipulation and the retesting intervention divided the sample into four groups: participants who were warned and flagged based on their high Time 1 Candidness scale score ($n = 60$), participants who were warned and were not flagged ($n = 54$), participants who were not warned and were flagged based on their high Time 1 Candidness scale score ($n = 77$), and participants who were not warned and were not flagged ($n = 44$).

**Results**

*Personality scales*

Means and standard deviations for both administrations of the Job Fit scales are presented in Table 1 by condition. The mean scores on the Job Fit scales for the flagged groups at Time 1 were higher than the mean scores for the no flag groups, indicating that the Candidness scale identified those individuals who may have been distorting their responses.

To evaluate the effects of the warning statement and retesting, a repeated-measures multivariate analysis of variance (MANOVA) was performed on the Job Fit scales. The results of the omnibus test are presented in Table 2. Significant multivariate effects were observed for
the Time and Flag main effects and the Time by Flag interaction. Univariate follow-up tests were conducted on each of the Job Fit scales and are presented in Table 3. Main effects for Time were observed for four of the six scales. In each case, mean scale scores at Time 1 were significantly higher than mean scores at Time 2. Main effects for Flag were found for each of the six scales. Examination of the means indicated that those individuals in the flagged groups scored significantly higher on each scale than those in the no flag groups. Interestingly, a main effect for Warn was observed for the Stress Tolerance scale; the mean Stress Tolerance score for individuals who did not receive the warning was significantly higher than the mean for those individuals who did receive the warning. The Time by Flag interaction was significant for each of the scales except the Assertiveness scale.

The nature of the Time by Flag interactions is important. If retesting is a viable policy, then individuals who have been flagged should have scores at Time 2 that are lower than their scores at Time 1. In contrast, those individuals who were not flagged should score similarly on both administrations of the test. Examination of the significant Time by Flag interactions indicated that the means were consistent with this pattern. As an example, Figure 1 presents a graph of this interaction for the Achievement Striving scale. In support of retesting, those in the flagged groups reduced their scores from the first to the second administration of the test while those in the no flag groups scored similarly on both administrations.

As an alternative way of exploring these interactions, standardized mean differences were calculated between Time 1 and Time 2 scores for each Job Fit scale within each condition. These effect sizes, which are presented in Table 4, characterize the extent to which the scores changed across the assessments; positive effect sizes indicate that scores at Time 2 were lower than scores at Time 1. The values for the flagged groups were positive and comparatively larger than those
for the no flag groups across both warning conditions. Time 2 scores for the participants in the
flagged groups decreased from 0.20 to as much as 1.07 standard deviation units relative to Time
1 scores. Comparatively, the no flag groups exhibited changes in their scale scores from -0.14 to
0.18 standard deviation units. Note that this pattern did not hold for the Assertiveness scale,
where scores actually increased slightly from Time 1 to Time 2, with larger increases found
among those who did not receive a warning. Comparing the effect sizes between the warning
conditions revealed an interesting pattern for the flagged groups. The values were consistently
smaller in the warning condition suggesting that the full effect of faking on the Job Fit scale
scores was reduced first through a warning, and then reduced further through retesting.

Because changes in scale means are not necessarily sensitive to rank-order changes
(Costa & McCrae, 1998; Holden, Helmes, Fekken, & Jackson, 1985), test-retest correlations
between Time 1 and Time 2 scores on the Job Fit scales were evaluated for each of the four
groups. These results are presented in Table 5. The test-retest correlations between Time 1 and
Time 2 for the no flag groups approximated normative test-retest reliability coefficients (Hastey
et al., 2001), whereas the test-retest correlations for the flagged groups were notably lower,
suggesting that red-flagging did change the positioning of the individuals. The differences
between the correlations computed for the flagged groups and the correlations for the no flag
groups were statistically significant for seven of the 12 comparisons made across the scales.

**Attitude Scales**

Means and standard deviations for both administrations of the attitude scales are also
presented in Table 1 by condition. The mean scores indicated that in general, levels of
respondent satisfaction, perceived fairness, test anxiety, and motivation decreased from Time 1
to Time 2, with few differential effects observed by retesting or warning condition. Interestingly,
the observed decrease in mean scores on the Test Anxiety scale is reflective of a positive impact on this respondent attitude.

To further evaluate the effects of the warning and retesting on respondent reactions, we conducted the same analyses as presented above on the attitude scales. The repeated-measures multivariate analysis of variance (MANOVA) on the attitude scales indicated significant multivariate main effects for Time and Flag (see Table 2). The univariate follow-up tests revealed main effects for Time on each of the four scales, with higher means observed for Time 1 (see Table 3). Main effects for Flag were found for perceived test fairness and test-taking motivation, with those in the flagged groups scoring higher on each measure. A main effect for Warn was found for test satisfaction, such that those in the warning condition reported that they were more satisfied with the test. The only significant interactions to emerge were Time by Warn and Time by Flag interactions for the Test Anxiety scale. Examination of the cell means indicated that both interactions were of the same form, reflecting a greater decrease in test anxiety scores between Time 1 and Time 2 for individuals who received a warning ($M_{Time\,1} = 22.14, M_{Time\,2} = 17.62$) relative to those who did not receive a warning ($M_{Time\,1} = 20.93, M_{Time\,2} = 18.16$) and for individuals in the no flag groups ($M_{Time\,1} = 22.78, M_{Time\,2} = 17.88$) relative to individuals in the flagged groups ($M_{Time\,1} = 20.62, M_{Time\,2} = 17.91$). A graphic representation of the latter interaction is provided as Figure 2.

The standardized mean difference effect sizes computed between Time 1 and Time 2 for each of the attitude scales were large and positive for all four groups (see Table 4). Scores on these scales decreased from Time 1 to Time 2 within each group. Specifically, Time 2 scores decreased from 0.42 to 0.80 standard deviation units relative to Time 1 scores for all conditions. Similarly, the test-retest correlations calculated between Time 1 and Time 2 for the attitude
scales were similar in value across the four conditions (see Table 5). Comparisons of the correlations between the flagged and no flag groups within each warning condition indicated no significant differences in the correlations. Thus, the respondents maintained a similar ordering in their attitudes across the assessments and the conditions.

Discussion

Although personality test scores have been shown to be good predictors of job performance, there is growing concern that socially desirable responding among applicants will negatively impact the usefulness of the scores. The inclusion of warning statements and the adoption of policies for retesting have both been offered as methods for dealing with this issue. In the present study, we evaluated these two alternatives.

Our results provide preliminary evidence that retesting is a viable method for managing faking tendencies. On the first administration of the test, individuals who were identified as engaging in socially desirable responding had mean personality scale scores that were notably higher than those who appeared to have answered in a less socially desirable manner. After retesting, scores for these two groups were much more similar. Examination of the effect sizes comparing personality scale scores at Time 1 to scores at Time 2 indicated that the flagged groups decreased their scores across the two assessments while scores for the no flag groups remained largely unchanged. Further, test-retest correlations computed between Time 1 and Time 2 scores suggested that retesting changed the relative orientation of the individuals in the flagged groups. Taken together, these results suggest that informing individuals, whose scores caused them to be red-flagged by an impression management scale, that they need to complete the test again, this time more honestly, can result in a new set of scores that are more comparable to those obtained by individuals who are less socially desirable in their responses.
Two factors are worthy of note when considering these results. First, scores for the flagged groups on the second administration were still higher than scores for the no flag groups. Although our data do not explain precisely why these individuals continued to score higher, we suspect that it is because they continued to engage in impression management during this administration, just to a lesser degree. These individuals may also have demonstrated “smarter” socially desirable responding. Within the flagged groups, the effect sizes for the Candidness scale between Time 1 and Time 2 were considerably larger than for any of the other scales. Thus, while the flagged participants did reduce their scores on the other scales considerably, they moderated their responses to the Candidness items to a greater extent.

Second, the general pattern of results regarding the effectiveness of retesting did not hold for the Assertiveness scale. This result was likely a product of the motivating instructions we employed. Our instructions were rather general, asking participants to respond as if they were applying for a desirable job. A high score on assertiveness may not have been viewed as more socially desirable in this ambiguous situation. Had we asked respondents to imagine a specific job in which being assertive was important, we expect that we would have found the same pattern of results as was observed for the other scales.

Contrary to previous research, we did not find supportive evidence for the use of warnings as a method for preventing faking tendencies (with the exception of the Stress Tolerance scale). This finding is not overly surprising in that previous research has noted that simply informing respondents that a faking scale is present does little to impact faking; the power of a warning comes in its pairing with certain consequences (Dwight & Donovan, 2003). Because we sought to experimentally manipulate the impact of retesting, we were unable to include the threat of consequences in our warning instructions. We speculate that if our warning
manipulation had been paired with the consequences necessary to produce a strong effect, a significant main effect for warning condition and a significant 3-way (Flag by Warn by Time) interaction would have emerged. The standardized mean difference values provided preliminary evidence that the techniques of warning and retesting may work together to reduce the impact of faking. The discovery of a significant 3-way interaction would indicate that individuals who are retested and warned demonstrate the greatest change in scores upon the second assessment. Thus, investigating the potential for preventative techniques to work in conjunction with correction techniques may represent a fertile ground for future research.

In addition to examining the effects of retesting and warnings on personality test scores, we also examined the effects on respondents’ reactions to the assessment experience. Retesting resulted in decreased levels of test satisfaction, perceived test fairness, test anxiety and test-taking motivation. Interestingly, it was not the experience of being told that they faked that produced the effect on these variables but the experience of taking the assessment again regardless of the purpose. While these changes in attitude scores were negative for the satisfaction, fairness, and motivation scales, the findings for test anxiety were positive. Participants reported less test anxiety at Time 2 than at Time 1, a result that suggests they became more comfortable with the assessment process when they had the opportunity to experience it for a second time.

Limitations

While our choice to investigate the impact of retesting through an experimental design was a logical first step, our use of a student sample and a laboratory setting is a clear limitation. Researchers have frequently cautioned against assuming that the results of directed-faking experiments will generalize to applicant contexts (Ellingson, et al., 2001; Hough, 1998; Rossé et
al., 1998; Smith, Hanges, & Dickson, 2001). For example, we observed an elevated distribution of Candidness scale scores relative to what one would observe in natural data. This extremity may explain why retesting decreased profile scale scores, but did not return the scores to a level commensurate with those individuals identified as responding honestly. Further, the fact that we uncovered attitude effects suggests that these effects may be even stronger in the field. The participants in this study made a personal commitment to devote the time necessary to take part in two sessions. In practice, individuals who are being assessed by organizations will not have made this type of commitment. Given that the stakes are clearly higher outside of the confines of the laboratory, it is imperative that these results be confirmed with organizational field data.

It is also important to recognize that we conducted our investigation using only one of the many personality measures that are available for commercial use. While the characteristics of the Job Fit made it an appropriate tool for this design, it is important that these results be generalized to other personality measures that contain faking scales. Differences between the Job Fit and these other measures could impact the effectiveness of retesting. Likewise, different impression management scales may measure faking in a different manner, and variance in the red-flagging process across personality scales could impact the extent to which retesting is effective. Also, the use of a measure with more than 60 items could further enhance the attitude effects observed as individuals are asked to devote more time and effort to completing the questions again. Therefore, replication of these results using other personality tools should further our understanding of retesting and its role as a corrective technique for managing faking tendencies.
Retesting

References


Footnotes

1 In this paper, we will use the term faking as synonymous with the terms impression management, response distortion, and socially desirable responding. Zerbe and Paulhus (1987) defined impression management as representing a deliberate misstatement of one’s characteristics. Thus, faking is treated in this paper as representing purposeful efforts to create a favorable impression.

2 The present criterion was established on the basis of normative data collected during the development of the Key Point Job Fit Assessment. The criterion was chosen so that individuals scoring above the criterion could reasonably be assumed to have engaged in socially desirable responding, yet not so high that it would have excluded excessively large numbers of respondents.
Table 1.

*Means and Standard Deviations for the Job Fit and Attitude Scales*

<table>
<thead>
<tr>
<th>Scale</th>
<th>Time 1</th>
<th>Time 2</th>
<th>Time 1</th>
<th>Time 2</th>
<th>Time 1</th>
<th>Time 2</th>
<th>Time 1</th>
<th>Time 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Achievement striving</td>
<td>34.4 (4.0)</td>
<td>32.6 (3.9)</td>
<td>30.4 (4.5)</td>
<td>29.7 (4.5)</td>
<td>35.6 (4.1)</td>
<td>33.0 (4.8)</td>
<td>30.4 (3.7)</td>
<td>30.7 (4.5)</td>
</tr>
<tr>
<td>Assertiveness</td>
<td>31.4 (4.6)</td>
<td>31.6 (4.9)</td>
<td>27.7 (4.9)</td>
<td>28.0 (5.4)</td>
<td>30.8 (4.4)</td>
<td>31.5 (4.4)</td>
<td>28.7 (5.0)</td>
<td>29.3 (5.6)</td>
</tr>
<tr>
<td>Dependability</td>
<td>35.7 (3.3)</td>
<td>34.5 (3.4)</td>
<td>33.3 (5.2)</td>
<td>32.9 (5.1)</td>
<td>36.7 (3.6)</td>
<td>35.0 (4.8)</td>
<td>33.7 (3.4)</td>
<td>33.3 (3.2)</td>
</tr>
<tr>
<td>Extroversion</td>
<td>32.3 (4.3)</td>
<td>31.3 (4.0)</td>
<td>28.9 (5.8)</td>
<td>29.2 (5.1)</td>
<td>32.4 (4.4)</td>
<td>31.1 (4.4)</td>
<td>28.8 (5.2)</td>
<td>28.5 (4.8)</td>
</tr>
<tr>
<td>Stress tolerance</td>
<td>31.4 (4.4)</td>
<td>30.3 (4.8)</td>
<td>28.5 (5.0)</td>
<td>29.1 (4.7)</td>
<td>33.1 (4.6)</td>
<td>32.4 (4.3)</td>
<td>29.9 (4.2)</td>
<td>30.0 (4.6)</td>
</tr>
<tr>
<td>Candidness</td>
<td>64.2 (3.5)</td>
<td>61.5 (3.9)</td>
<td>56.4 (3.9)</td>
<td>56.7 (4.0)</td>
<td>65.5 (3.8)</td>
<td>61.9 (4.6)</td>
<td>56.7 (4.4)</td>
<td>57.0 (4.5)</td>
</tr>
<tr>
<td>Test satisfaction</td>
<td>10.8 (2.1)</td>
<td>9.4 (2.5)</td>
<td>10.4 (2.2)</td>
<td>8.9 (2.3)</td>
<td>10.3 (2.5)</td>
<td>8.8 (2.6)</td>
<td>9.8 (2.1)</td>
<td>8.3 (2.3)</td>
</tr>
<tr>
<td>Perceived test fairness</td>
<td>16.9 (3.7)</td>
<td>14.6 (4.6)</td>
<td>15.1 (3.5)</td>
<td>13.5 (4.1)</td>
<td>15.5 (4.3)</td>
<td>13.9 (4.6)</td>
<td>15.2 (4.4)</td>
<td>12.8 (4.6)</td>
</tr>
<tr>
<td>Test anxiety</td>
<td>21.5 (7.2)</td>
<td>18.4 (7.3)</td>
<td>22.9 (7.2)</td>
<td>16.8 (4.6)</td>
<td>20.0 (5.3)</td>
<td>17.6 (4.9)</td>
<td>22.6 (6.5)</td>
<td>19.2 (7.0)</td>
</tr>
<tr>
<td>Test-taking motivation</td>
<td>25.3 (3.8)</td>
<td>23.4 (5.1)</td>
<td>23.6 (3.9)</td>
<td>21.0 (5.2)</td>
<td>24.9 (4.1)</td>
<td>23.4 (5.2)</td>
<td>22.9 (3.6)</td>
<td>20.7 (5.5)</td>
</tr>
</tbody>
</table>
Table 2.

Repeated-measures MANOVA for the Job Fit and Attitude Scales

<table>
<thead>
<tr>
<th>Source</th>
<th>Job Fit scales</th>
<th></th>
<th>Attitude scales</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>( F )</td>
<td>( p )</td>
<td>( F )</td>
</tr>
<tr>
<td><strong>Within-subjects effects</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>8.27</td>
<td>.00</td>
<td>73.21</td>
<td>.00</td>
</tr>
<tr>
<td>Time x Warn</td>
<td>0.57</td>
<td>.75</td>
<td>1.41</td>
<td>.23</td>
</tr>
<tr>
<td>Time x Flag</td>
<td>8.63</td>
<td>.00</td>
<td>2.17</td>
<td>.07</td>
</tr>
<tr>
<td>Time x Warn x Flag</td>
<td>1.23</td>
<td>.29</td>
<td>1.49</td>
<td>.21</td>
</tr>
<tr>
<td><strong>Between-subjects effects</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Warn</td>
<td>1.80</td>
<td>.10</td>
<td>1.42</td>
<td>.23</td>
</tr>
<tr>
<td>Flag</td>
<td>31.67</td>
<td>.00</td>
<td>5.86</td>
<td>.00</td>
</tr>
<tr>
<td>Warn x Flag</td>
<td>0.48</td>
<td>.82</td>
<td>1.21</td>
<td>.31</td>
</tr>
</tbody>
</table>

*Note.* Time = repeated measurement from Time 1 to Time 2, Warn = warning condition, Flag = flagging condition. MANOVA df (6, 228).
Table 3.

F-values from the Univariate Tests for the Job Fit and Attitude Scales

<table>
<thead>
<tr>
<th>Scale</th>
<th>Within-subjects effects</th>
<th>Between-subjects effects</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Time</td>
<td>Time x Warn</td>
</tr>
<tr>
<td>Achievement striving</td>
<td>25.27**</td>
<td>0.03</td>
</tr>
<tr>
<td>Assertiveness</td>
<td>3.35</td>
<td>0.84</td>
</tr>
<tr>
<td>Dependability</td>
<td>17.76**</td>
<td>0.41</td>
</tr>
<tr>
<td>Extroversion</td>
<td>5.91*</td>
<td>0.85</td>
</tr>
<tr>
<td>Stress tolerance</td>
<td>1.13</td>
<td>0.01</td>
</tr>
<tr>
<td>Candidness</td>
<td>31.98**</td>
<td>0.77</td>
</tr>
<tr>
<td>Test satisfaction</td>
<td>121.91**</td>
<td>0.07</td>
</tr>
<tr>
<td>Perceived test fairness</td>
<td>91.57**</td>
<td>0.02</td>
</tr>
<tr>
<td>Test anxiety</td>
<td>107.73**</td>
<td>5.47*</td>
</tr>
<tr>
<td>Test-taking motivation</td>
<td>75.98**</td>
<td>0.71</td>
</tr>
</tbody>
</table>

Note. Time = repeated measurement from Time 1 to Time 2, Warn = warning condition, Flag = flagging condition.

* F-value significant at $p < .05$

** F-value significant at $p < .01$
Table 4.

**Standardized Group Differences Computed between Time 1 and Time 2 Scores for the Job Fit and Attitude Scales**

<table>
<thead>
<tr>
<th>Scale</th>
<th>Warning</th>
<th></th>
<th></th>
<th>No warning</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Flagged</td>
<td>No flag</td>
<td>Flagged</td>
<td>No flag</td>
<td>Flagged</td>
<td>No flag</td>
</tr>
<tr>
<td>Achievement striving</td>
<td>.52</td>
<td>.18</td>
<td>.73</td>
<td>-.08</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assertiveness</td>
<td>-.03</td>
<td>-.07</td>
<td>-.19</td>
<td>-.13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dependability</td>
<td>.45</td>
<td>.08</td>
<td>.57</td>
<td>.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extroversion</td>
<td>.29</td>
<td>-.07</td>
<td>.34</td>
<td>.07</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stress tolerance</td>
<td>.30</td>
<td>-.14</td>
<td>.20</td>
<td>-.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Candidness</td>
<td>.88</td>
<td>-.09</td>
<td>1.07</td>
<td>-.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test satisfaction</td>
<td>.75</td>
<td>.80</td>
<td>.67</td>
<td>.79</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perceived test fairness</td>
<td>.67</td>
<td>.49</td>
<td>.42</td>
<td>.57</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test anxiety</td>
<td>.47</td>
<td>.92</td>
<td>.49</td>
<td>.57</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test-taking motivation</td>
<td>.58</td>
<td>.76</td>
<td>.42</td>
<td>.70</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note. d = standardized group difference. Positive values indicate a decrease in scores from Time 1 to Time 2.*
Table 5.

Test-retest Correlations and Significant Difference Tests Comparing the Test-retest Correlations for the Job Fit and Attitude Scales

<table>
<thead>
<tr>
<th>Scale</th>
<th>Warning Flagged</th>
<th>Warning No flag</th>
<th>No warning Flagged</th>
<th>No warning No flag</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$r_{12}$</td>
<td>$r_{12}$</td>
<td>$z$</td>
<td>$r_{12}$</td>
</tr>
<tr>
<td>Achievement striving</td>
<td>.64</td>
<td>.84</td>
<td>2.40*</td>
<td>.45</td>
</tr>
<tr>
<td>Assertiveness</td>
<td>.72</td>
<td>.80</td>
<td>.99</td>
<td>.60</td>
</tr>
<tr>
<td>Dependability</td>
<td>.73</td>
<td>.83</td>
<td>1.34</td>
<td>.51</td>
</tr>
<tr>
<td>Extroversion</td>
<td>.64</td>
<td>.83</td>
<td>2.23*</td>
<td>.61</td>
</tr>
<tr>
<td>Stress tolerance</td>
<td>.62</td>
<td>.77</td>
<td>1.53</td>
<td>.43</td>
</tr>
<tr>
<td>Candidness</td>
<td>.66</td>
<td>.74</td>
<td>0.81</td>
<td>.31</td>
</tr>
<tr>
<td>Test satisfaction</td>
<td>.62</td>
<td>.75</td>
<td>1.29</td>
<td>.59</td>
</tr>
<tr>
<td>Perceived test fairness</td>
<td>.71</td>
<td>.74</td>
<td>0.33</td>
<td>.74</td>
</tr>
<tr>
<td>Test anxiety</td>
<td>.70</td>
<td>.57</td>
<td>-1.14</td>
<td>.53</td>
</tr>
<tr>
<td>Test-taking motivation</td>
<td>.71</td>
<td>.75</td>
<td>0.45</td>
<td>.75</td>
</tr>
</tbody>
</table>

Note. $r_{12}$ = test-retest correlation, $z$ = z-value for tests for significant differences between independent correlations. All correlations are significant at $p < .01$.

* z-value significant at $p < .05$

** z-value significant at $p < .01$
Figure Captions

*Figure 1.* The significant Time by Flag interaction observed for the Achievement Striving scale.

*Figure 2.* The significant Time by Flag interaction observed for the Test Anxiety scale.
Achievement Striving scale means

- Flagged
- No flag