Impact of electronic warnings on online personality scores and test-taker reactions in an applicant simulation

Gary N. Burns *, Jenna N. Fillipowski, Megan B. Morris, Elizabeth A. Shoda
Wright State University, 3640 Colonel Glenn Hwy, Dayton, OH 45435, United States

**Article info**

**Article history:**
Available online 16 February 2015

**Keywords:**
Personality
Warnings
Computer assessment
Faking
Assessment

**Abstract**

This study investigates the impact of different types of electronic warnings during a simulated job application assessment. Results indicated that negatively worded warnings and accusations worked better than positively worded warnings at blunting faking. Although there was some evidence that test-takers engaging in higher levels of faking heeded warnings more, warnings tended to decrease scores for all test-takers. While positive warnings motivated test-takers to perform well, negative warnings and accusations increased test-taking anxiety. Whereas past research has failed to find a relationship between warnings and perceived fairness, current results suggest that this relationship depends on the level of test-taker distortion.

© 2015 Elsevier Ltd. All rights reserved.

**1. Introduction**

Despite the popular use of personality inventories, a concern with applied personality measurement is that job applicants will distort, or fake, their test behavior, resulting in an overly-positive score. Research widely supports this fear, with estimates indicating that somewhere between a quarter to half of surveyed job applicants distort their responses to personality measures (Donovan, Dwight, & Hurtz, 2003; Griffith, Chmielowski, & Yoshita, 2007; Landers, Sackett, & Tuzinski, 2011). While mixed results have led some to question if this type of faking is a concern (Hogan, Barrett, & Hogan, 2007; Hough, 1998; Ones, Viswesvaran, & Reiss, 1996), the evidence that faking is a threat to hiring decisions and construct validity is becoming more difficult to dismiss (Marcus, 2006; Mueller-Hanson, Heggstad, & Thornton, 2003; Rosse, Stecher, Miller, & Levin, 1998).

The assertion that faking might cause adverse outcomes is supported by the argument that when selection ratios are low, people who fake have a disproportionately greater chance of being hired, even if the criterion-related validity is unaffected (Marcus, 2006). Several studies have illustrated that faking can also have implications for hiring decisions (Christiansen, Goffin, Johnston, & Rothstein, 1994). Furthermore, claims that faking has no impact on predictive validity are weakened by the fact that scales designed to measure faking (e.g., social desirability) are correlated to honest personality scores of interest (e.g., conscientiousness, agreeableness, and emotional stability) (McCrae & Costa, 1983; Ones et al., 1996). This makes it difficult to tell whether applicants who score high on social desirability scales are faking, if they truly have elevated personality traits (Goffin & Christiansen, 2003), thus giving less credibility to the viewpoint that faking has little to no impact on criterion-related validity, and reigniting an interest in further exploring the antecedents and outcomes of response distortion.

The concern about faking has inspired a wide range of techniques to counteract applicant distortion. These techniques can generally be broken down into two broad categories, proactive or reactive, and a third that combines both. First, there are proactive attempts to keep applicants from engaging in faking behavior. Such methods include the use of social desirability scores as a proxy of faking behavior (Burns & Christiansen, 2006; Peterson, Griffith, Isaacson, O’Connell, & Mangos, 2011), but other techniques might have promise (i.e., Bing et al.’s, 2011).
Of the proactive strategies, the easiest one to implement with existing procedures is warnings. Unfortunately, researchers have also expressed concerns about warnings. Several researchers have cautioned that warnings might negatively impact test-taker reactions (Tett et al., 2006; Zickar & Gibby, 2006). Tett et al. (2006) also cautioned that warnings might cue risk-takers to attempt faking whereas Zickar and Gibby (2006) warned that, over time, test-takers would become aware of this method and the limited capability to detect fakers accurately. These concerns led Tett et al. (2006) to recommend that these warnings might be best used as a supplement to other types of faking detection measures. Taking advantage of such pairing, it is possible to target warnings only to those test-takers that are believed to be engaging in faking behavior (see Fan et al., 2012; Landers et al., 2011).

A third technique combines proactive and reactive strategies to target interventions to test-takers that are believed to be engaging in faking behavior (see Landers et al., 2011). Fan et al. (2012) illustrates the impact of specifically targeting such electronic warnings to test-takers identified early in the testing process. While such targeted warnings can easily be delivered, our current understanding and measurement of faking limits the accuracy in the administration of these warnings. While research on alternative faking measures continues to develop, the large body of extant research indicates that such electronic warnings will be rife with false positive and false negatives. As such, it is necessary to understand the impact of electronic warnings at multiple levels of various indicators of distortion. In an attempt to increase knowledge in this area, we explored the use of three different types of electronically delivered warnings in a laboratory experiment. Taking advantage of the laboratory setting we were able to conduct a true experiment to examine the impact of warnings across participants engaging in different levels of faking without ethically compromising personnel processes.

1.1. Effects of warnings on test scores

Many studies have shown the effectiveness of introducing a warning statement during a personality test administration, with Landers et al. (2011) providing strong evidence that warnings reduced the amount of blatant extreme responding in real world settings. In a meta-analytic review of the warning literature, Dwight and Donovan (2003) examined the effectiveness of warning applicants not to fake on personality measures among different types of warnings. Overall, test-takers who received warnings had test scores that were .23 standard deviations lower on average than unwarned test-takers. They also found that the type of warning moderated the efficacy of the warnings. Warnings regarding identification (e.g., this test contains items to identify fakers) had very little effect on test-taking behavior, with only a .01 standard deviation difference between groups. Consequence based warnings (e.g., dishonesty will not be tolerated) were the most effective type of warning, reducing warned test-takers’ scores by .30 standard deviations.

Whereas most warnings used in published research have been mild, there is also the capability of accusing applicants of faking (e.g., Landers et al., 2011). Although such an accusation is an extreme reaction, its use is not unprecedented in the field of personality testing. For example, it is recommended that practitioners interpreting MMPI profiles marked as invalid (i.e., a profile with an elevated L, K, or F scale) refrain from making interpretation conclusions (Butcher, 1977). One recommended solution is to inform applicants that their scores are uninterpretable and to have them complete the test again (Butcher, Morfitt, Rouse, & Holden, 1997; Cigrang & Staal, 2001). Whereas researchers with the MMPI have primarily focused on retesting due to defensiveness, Ellingson, Heggestad, and Makarius (2012) found that similar effects were observed when asking test-takers to retest due to intentional distortion. These results indicate that when applicants are accused of distorting their responses, a second test administration typically results in more accurate personality scores. It should be noted that while research on these types of accusations is rare, that this research has emerged indicates that some organizations have already begun exploring such procedures.

1.2. Potential negative effects of warnings

Although warnings might offer a method of controlling or suppressing faking behavior, there are two potential negative consequences of warnings. First, warnings might have a negative effect on test-takers’ reactions to the testing process. Despite these concerns, warnings appear to have very little effect on test-takers’ reactions and attitudes toward employment screening measures that measure personality traits. Examining procedural justice perceptions, McFarland (2003) found no significant difference between warned and unwarned test-takers. Similarly, Converse et al. (2008) found that neither positive nor negative warnings influenced perceived test ease, test-taker expectations, satisfaction with the testing process, or belief in the tests. However, Converse et al. (2008) did find that the negative warnings resulted in a higher level of test-taker anxiety. In our study, we wanted to focus on the effects of an extreme negative warning, an accusation of faking, on test anxiety and perceptions of fairness.

Second, warnings might also distort the nature of the personality data gathered. Robson, Jones, and Abraham (2008) found that warning statements reduced the convergent validity of the personality dimensions. This is also consistent with results reported by McFarland (2003), who found that correlations among personality variables were stronger in unwarned conditions compared to warned conditions. Perhaps most concerning, recent experiments by Fan et al. (2012) and Ellingson et al. (2012) indicate that warnings or retesting could cause non-faking test-takers to reduce their scores, especially when these individuals have been falsely identified as distorting their responses on computer administered tests. Both Fan et al. (2012) and Ellingson et al. (2012) provide convincing evidence that when presented with warnings of intentional distortion, even test-takers not actively distorting their responses will reduce their scores. Although such applicant distortion is different than the phenomenon of faking, it is a form of response distortion that practitioners should be concerned with.

Based on these results, practitioners should carefully consider the potential trade-offs of using warning statements. Strong evidence suggests that blatant distortion will be reduced (Landers et al., 2011) and the overall impact of the situational pressure to fake will be blunted (Dwight & Donovan, 2003); however, this might not influence criterion related validity coefficients (Converse et al., 2008). Furthermore, use of these statements might also make test users more anxious and might result in the degradation of the construct validity of the personality measures (Robson et al., 2008; Vasilopoulos, Cucina, & McElreath, 2005) and inaccurate scores for honest test-takers (Ellingson et al., 2012; Fan et al., 2012). One goal of the current study is to examine how these warnings influence responses to personality screening instruments across test takers responding in different ways.

1.3. Effects of warning at different levels of distortion

Additionally, warnings might not be effective for all respondents. Tett et al. (2006) cautions that warnings might cue risk-takers to engage in faking behavior, whereas other applicants might not believe the message. In addition to these concerns, there is also reason to believe that warnings will have differential effects on applicants engaging in different levels of faking.
As discussed above, not all applicants engage in faking and not all fakers engage in distortion equally (McFarland & Ryan, 2006). Beyond limitations imposed by ceiling effects, research has revealed differences in faking extremity among both job applicants (Birkeland, Manson, Kisamore, Brannick, & Smith, 2006) and participants in simulated applicant conditions (Viswesvaran & Ones, 1999). We anticipate that warnings will have a differential impact on test-takers reactions and behavior depending on the level of distortion that they are engaging in. Focusing first on test-taking behavior, we expect that test-takers engaging in more extreme levels of faking will more drastically alter their responses after receiving an electronically delivered warning than those applicants engaging in low or moderate levels of faking. Such effects might simply be the effect of extreme fakers having more room to decrease their scores following warnings; such a pattern would reflect the goal of using warnings. Alternatively, if warnings cause similar decreases across levels of distortion this might be concerning, as such a uniform decrease would indicate distortion in the personality constructs being measured, not necessarily a reduction in faking.

In addition, these warnings might also convey information to test-takers about the fairness of the testing procedure. Test-takers not engaging in faking should see these warnings as the organization’s commitment to the procedural justice of the testing process and their attempt to keep a level playing field for all applicants. However, those who are engaging in more extreme levels of faking should be disturbed by warnings that there are techniques to detect honest responses.

2. Current study

The majority of the research on faking has examined the impact of warnings delivered before test-takers begin the test-taking process. To date, research indicates that such warnings do cause meaningful score differences between test-takers who receive different warnings or no warnings. More recent research on warnings has turned to examining the ability of researchers to deliver warnings electronically during the actual testing process to targeted individuals (e.g., Landers et al., 2011). Such procedures could also be developed based on other measures of faking behavior, such as social desirability scales or overclaiming measures (Bing et al., 2011). Although such messages are usually explored as warnings with consequences, research by Butcher et al. (1997), Cigrang and Staal (2001), and Ellingson et al. (2012) indicate that some organizations are already implementing accusations as a form of warning.

The purpose of the current study was to evaluate the use of electronically delivered warnings and accusations on test-taking behavior. We sought to deliver both positively and negatively worded warnings during an employment screening measure to see how individual test-takers changed their responses after receiving the warnings. Based on the research reviewed above, we hypothesized that these warnings would suppress personality scores in a group responding under job application instructions (Hypothesis 1). We also wanted to introduce a stronger warning more similar to the accusations presented by Butcher et al. (1997), Cigrang and Staal (2001), and Ellingson et al. (2012). We hypothesized that accusations of faking would suppress scores more than either positive or negatively worded warnings (Hypothesis 2).

Based on the results of Converse et al. (2008) we also developed specific hypotheses about the impact of warnings on test-takers reactions. Attempting to replicate Converse et al.’s results, we hypothesized that negative warnings and accusations would lead to increased test anxiety (Hypothesis 3). Also, because of their extreme wording, we believed that the accusations would reduce test-takers perceptions of the fairness and accuracy of the testing process (Hypothesis 4). Finally, it is possible that these warnings will have a positive effect on test-taker motivation. Specifically, positively worded warning messages might remind test-takers of the importance of their results and motivate them to do well on the test without distorting their responses (Hypothesis 5).

Because we generally do not believe that warnings offer a blanket solution for all test-takers, we developed several hypotheses examining the moderating effects of warnings across applicants. First, we hypothesized that warnings would have a greater effect on personality scores for test-takers’ engaging in extreme levels of faking compared to those engaging in no to moderate levels of faking (Hypothesis 6). Second, we hypothesized that those applicants engaging in no to moderate levels of faking will have more positive perceptions of fairness and accuracy of the testing process after receiving positive or negatively worded warnings than those engaging in extreme levels of faking (Hypothesis 7), though this effect will likely reverse in the accusation condition (Hypothesis 8). Although not directly expressed above, the rationale is that test-takers not engaging in faking will view such warnings as a leveling of the testing environment while those engaging in faking will view such warnings as impediment to their adopted test taking strategy.

3. Method

3.1. Participants

Four hundred and fifty-seven undergraduate students were recruited from an undergraduate participant pool and offered course credit for participation. Sixty-five percent of the sample was female, with an average age of 19.77 years (SD = .66). The majority of participants (66%) self-identified as Caucasian with the rest identifying themselves as African American (22%), Hispanic (3%), Asian (3%), or multiracial/other (6%).

3.2. Procedure

Students were randomly assigned into one of five different conditions: one honest, baseline, condition and four job applicant conditions. Assignment to the honest, baseline condition was staggered to produce a smaller sample size for this condition, as its sole purpose was to serve as a faking manipulation check. In this honest condition participants were told to answer the questions as honestly as possible. Recruitment continued in the job applicant conditions until there were 100 participants in each condition to allow for equal sample sizes across the conditions of interest.

For both the honest and job applicant conditions the testing process was the same. In the job applicant conditions, participants were provided with a job description and told that they were helping to test a program for a local organization hiring customer service and sales representatives. Participants were told that test scores would typically be used to hire associates, but because this was an applicant simulation, the participants “hired” would receive $50, not a real job offer. Participants were not told the number of “job openings” for the position. The test battery consisted of four components. Part 1 was a job experience section containing bogus items; bogus items reflect work tasks that do not exist and offer a more objective measure of faking behavior (Burns & Christiansen, 2011). Part 2 was a verbal ability test. Part 3 was a quantitative ability test. Part 4 was the employment screening measure that contained the experimental induction. Within the job applicant condition, participants were randomly assigned to one of four warning conditions: a no-warning progress report, a positive warning, a negative warning, and an accusation. Warnings were electronically delivered after the participants...
completed 25 items in the personality inventory. In the no-warning and baseline honest conditions, participants received a progress report informing them that the testing process was almost done:

Thank you for participating in this study. You are almost done with the application process.

All other warning statements were prefaced with the following statements:

It is critical to note that these inventories have items which are designed to detect faking. Research has shown that these questions are able to identify individuals who provide inaccurate information about themselves.

In the positive warning condition participants also received the additional statement:

If you answer honestly, then you will be qualified to win $50. In other words, you get a chance to earn some money if you answer accurately and honestly as possible.

In the negative warning condition participants received the additional statement:

If you answer dishonestly, then you will be automatically disqualified from winning the $50. In other words, you will lose the chance to earn money if you answer inaccurately or dishonestly.

The accusation condition included the additional statement:

You have been flagged as being dishonest. If you continue to answer dishonestly, then you will be automatically disqualified from winning the $50.

Finally, after completing the four components of the selection process, participants completed a survey examining test-taker reactions. To distance these ratings from the test responses, participants were told that the testing procedure was finished and that their answers to this section would not be used to determine who was “hired” and that this information would be of great value to the local organization.

3.3. Measures

3.3.1. Pre-warning personality

Before the warning was delivered, participants completed Donnellan, Oswald, Baird, and Lucas’s (2006) 20-item IPIP measure of the Big Five. Each Big Five factor was measured with four items. Participants responded on a 5-point scale ranging from 1, “Strongly Disagree”, to 5, “Strongly Agree”. Cronbach reliability estimates were .73 for Extraversion, .68 for Agreeableness, .76 for Conscientiousness, .64 for Emotional Stability, and .65 for Openness. Scores are keyed so that higher scores indicate more desirable responses.

3.3.2. Post-warning personality

After delivery of the warnings participants completed the 50-item IPIP measures of the Big Five, which also contained the 20 Donnellan et al. (2006) items. Each Big Five factor was measured with 10 items. Participants again responded on a 5-point scale and Cronbach reliability estimates were .90 for Extraversion, .84 for Agreeableness, .73 for Conscientiousness, .87 for Emotional Stability, and .81 for Openness. Scores are keyed so that higher scores indicate more desirable responses.

3.3.3. Pre-warning impression management

Impression management scales are one of the most common methods of estimating applicant faking and flagging participants’ responses (Burns & Christiansen, 2006). Impression management was assessed with five items from the Balanced Inventory of Desirable Responding (BIDR: Paulhus, 1998) as part of the pre-warning personality items. Participants rated items on the same 5-point scale as the personality items. Participants received one point when providing an extreme response of “Strongly Disagree” or “Strongly Agree” to an impression management item, such that higher scores represent more extreme socially desirable responding. Cronbach’s reliability estimate for the five items was .61. Selection of these five items was based on factor loadings and mean-differences from a previous sample examining the same population and job applicant manipulation.

3.3.4. Post-warning impression management

After delivery of the warnings participants completed the 20-item Impression Management scale from the BIDR (Paulhus, 1998). Participants again responded on a 5-point scale and the Cronbach reliability estimate was .88. The short, pre-progress report IM scores were correlated .72 and .70 with the full, post-message IM scores in the honest baseline and no-warning conditions.

3.3.5. Bogus items

Bogus items represent another method of identifying applicants engaging in faking behavior (Anderson, Warner, & Spencer, 1984; Burns & Christiansen, 2011). In the first part of the assessment participants completed a 12-item experience questionnaire. Participants were asked to indicate that “Yes” they had experience with the task, “No” they did not have experience with the task, or “Maybe” they had experience with the task. Two example experience items are: “I have experience making cold calls” and “I have experiencing qualifying sales prospects.” Five of the experience items were bogus items (Anderson et al., 1984), representing non-existing tasks. Two example bogus items are: “I have experience with the alpha and beta call techniques” and “I am aware of the American Marketing Organization’s Salesperson’s Bill of Rights”. Participants received one point when responding “Yes” to a bogus item, with high scores indicating more pre-warning distortion on the experience section. The Cronbach’s reliability estimate for the five items was .69. Items were taken from Tristan (2009), who demonstrated that applicants scored higher on these items than incumbents.

3.3.6. Applicant reactions

Test-taker motivation was measured with four items (Cronbach’s alpha = .78) from Arvey, Strickland, Drauden, and Martin (1990). An example item is “I want to be among the top scorers on these tests.” Test anxiety was measured with five items (α = .75) from Arvey et al. (1990). An example item is “During the testing, I got so nervous I could not do as well as I should have.” Face validity, or perceived accuracy, was measured with five items (α = .86) from Smither, Reilly, Millsap, Pearlman, and Stoffey (1993). An example item is “I could not see any relationship between the examination and what is required on the job.” Procedural justice, or perceived fairness, was measured with two items (α = .77) from Smither et al. (1993). An example item is “Overall, I believe that the examination was fair.”

4. Results

For ease of interpretation we present the impact of warnings on test scores (Hypotheses 1, 2, and 6) and applicant reactions (Hypotheses 3, 4, 5, 7, and 8) grouped separately. Because of the
large number of variables examined across four conditions and the specific nature of our hypotheses, no summary correlation matrices are included – interested readers are encouraged to contact the first author for these. Means and standard deviations for study variables can be found in Table 1.

### 4.1. Applicant instructions manipulation check

We first examined pre-warning scores across the applicant simulation groups to identify if there were any initial significant differences between the applicant instructions control group and the three warning groups. Using all Big Five variables as well as the impression management scale and the bogus item scale, results of a MANOVA indicated that there were no significant differences in pre-warning scores between the applicant instructions conditions (Wilks’ Lambda = .95, p > .05). These results indicate that participants in the job applicant conditions were responding similarly before the warning. Collapsing across applicant instruction conditions, we repeated our MANOVA to examine differences in pre-warning scores between the aggregated applicant instructions group and an honest control group. Results of this MANOVA did indicate significant differences (Wilks’ Lambda = .91, p < .05), with follow-up analyses indicating that scores were significantly higher in the job applicant conditions than in the baseline honest condition. Specifically, Cohen’s d indicated that the applicant group scored .31 SDs higher on the impression management scale, .75 SDs higher on the bogus item scale, .64 SDs higher on the extraversion scale, .41 SDs higher on the agreeableness scale, .36 SDs higher on the conscientiousness scale, .19 SDs higher on the emotional stability scale, and .69 SDs higher on the openness scale. Together, these results indicate that our applicant instructions induced faking and that there were no initial significant differences between applicant instruction groups prior to receiving warnings.

### 4.2. Effects of Warnings on Test Scores

Each post-warning personality measure was analyzed separately using analysis of covariance (ANCOVA) to compare the post-warning scores across conditions. ANCOVA was chosen in order to adjust the post-warning means to account for pre-warning differences and to increase the power of the analyses by reducing within-group variability (Huck & McLean, 1975). Arvey and Cole (1989) have shown the ANCOVA procedures to be more powerful than the posttest-only or difference score designs when evaluating change in organizational settings (see also Kisbu-Sakarya, MacKinnon, & Aiken, 2012).

Each job applicant group was constrained to have equal sample sizes to allow for use of Type 1 sum of squares. Type 1 sum of squares was chosen so that confounded variance in the ANCOVA sizes to allow for use of Type 1 sum of squares. Type 1 sum of squares was chosen so that confounded variance in the ANCOVA measure and then to the condition. This procedure was selected because we believed that the best estimate of post-warning scores would be pre-warning scores. Table 1 presents the unadjusted means for the four applicant conditions as well as the unadjusted baseline honest scores for comparison purposes. Fig. 1 displays the adjusted means and standard errors for all five personality traits across applicant conditions.

Hypothesis 1, that different types of warnings would result in different scores, was tested by specifying an ANCOVA model with the pre-warning score as a covariate and the experimental warning condition as the between-subject factor. Results indicated that post-warning scores varied across warning conditions for extraversion (F(3, 395) = 5.94, p < .05, partial $\eta^2 = .04$), agreeableness (F(3, 395) = 6.39, p < .05, partial $\eta^2 = .05$), conscientiousness (F(3, 395) = 10.57, p < .05, partial $\eta^2 = .07$), emotional stability (F(3, 395) = 4.84, p < .05, partial $\eta^2 = .04$), and openness (F(3, 395) = 6.02, p < .05, partial $\eta^2 = .04$). The comparison of positive warnings, negative warnings, and accusations to the no-warning group is presented below, stratified by type of warning. Using a Bonferroni correction to account for family-wise error rates across traits, the modified alpha for Hypothesis 1 and 2 was .01.

#### 4.2.1. Impact of positively worded warnings

Controlling for pre-warning scores, post-warning scores were significantly lower for openness ($t(198) = -2.46$, p < .05) in the positive warning condition than in the no-warning condition. No significant differences were present for extraversion ($t(198) = -1.70$, p > .01), agreeableness ($t(198) = -.40$, p > .01), conscientiousness ($t(198) = -.80$, p > .05), and emotional stability ($t(198) = -1.71$, p > .01). Although means were in the expected direction, this pattern of results provides limited support for Hypothesis 1 regarding positive warnings.

### Table 1

Unadjusted means across conditions.

<table>
<thead>
<tr>
<th>Scale</th>
<th>Honest baseline</th>
<th>No warning control</th>
<th>Positive warning</th>
<th>Negative warning</th>
<th>Accusation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Impression management</strong></td>
<td>1.35 (1.29)</td>
<td>1.55 (1.31)</td>
<td>1.83 (1.23)</td>
<td>1.80 (1.41)</td>
<td>2.02 (1.63)</td>
</tr>
<tr>
<td>Bogus items</td>
<td>1.04 (1.15)</td>
<td>2.01 (1.53)</td>
<td>2.12 (1.62)</td>
<td>2.02 (1.49)</td>
<td>2.16 (1.68)</td>
</tr>
<tr>
<td>Extraversion</td>
<td>3.45 (0.67)</td>
<td>3.86 (0.70)</td>
<td>3.87 (0.72)</td>
<td>3.95 (0.59)</td>
<td>3.76 (0.75)</td>
</tr>
<tr>
<td>Agreeableness</td>
<td>3.93 (0.62)</td>
<td>4.17 (0.57)</td>
<td>4.21 (0.53)</td>
<td>4.20 (0.59)</td>
<td>4.22 (0.35)</td>
</tr>
<tr>
<td>Conscientiousness</td>
<td>3.96 (0.69)</td>
<td>4.24 (0.61)</td>
<td>4.28 (0.63)</td>
<td>4.32 (0.59)</td>
<td>4.29 (0.63)</td>
</tr>
<tr>
<td>Emotional stability</td>
<td>3.65 (0.66)</td>
<td>3.72 (0.70)</td>
<td>3.85 (0.69)</td>
<td>3.78 (0.64)</td>
<td>3.77 (0.67)</td>
</tr>
<tr>
<td>Openness</td>
<td>3.68 (0.60)</td>
<td>4.12 (0.55)</td>
<td>4.10 (0.56)</td>
<td>4.04 (0.60)</td>
<td>4.09 (0.58)</td>
</tr>
<tr>
<td><strong>Post-warning scores</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Impression management</td>
<td>2.79 (4.05)</td>
<td>4.37 (4.92)</td>
<td>3.65 (3.94)</td>
<td>3.41 (3.60)</td>
<td>3.21 (3.81)</td>
</tr>
<tr>
<td>Extraversion</td>
<td>3.38 (0.61)</td>
<td>3.80 (0.68)</td>
<td>3.72 (0.71)</td>
<td>3.74 (0.65)</td>
<td>3.49 (0.71)</td>
</tr>
<tr>
<td>Agreeableness</td>
<td>3.84 (0.55)</td>
<td>4.09 (0.54)</td>
<td>4.11 (0.49)</td>
<td>4.08 (0.45)</td>
<td>3.96 (0.48)</td>
</tr>
<tr>
<td>Conscientiousness</td>
<td>3.67 (0.49)</td>
<td>3.86 (0.47)</td>
<td>3.84 (0.42)</td>
<td>3.78 (0.39)</td>
<td>3.67 (0.43)</td>
</tr>
<tr>
<td>Emotional stability</td>
<td>3.67 (0.65)</td>
<td>3.75 (0.67)</td>
<td>3.75 (0.61)</td>
<td>3.75 (0.59)</td>
<td>3.58 (0.61)</td>
</tr>
<tr>
<td>Openness</td>
<td>3.58 (0.37)</td>
<td>3.94 (0.53)</td>
<td>3.80 (0.58)</td>
<td>3.76 (0.52)</td>
<td>3.71 (0.53)</td>
</tr>
<tr>
<td><strong>Test takers reactions</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Motivation</td>
<td>–</td>
<td>3.89 (0.71)</td>
<td>4.08 (0.59)</td>
<td>3.95 (0.62)</td>
<td>3.94 (0.62)</td>
</tr>
<tr>
<td>Test anxiety</td>
<td>–</td>
<td>2.17 (0.64)</td>
<td>2.21 (0.70)</td>
<td>2.34 (0.79)</td>
<td>2.33 (0.61)</td>
</tr>
<tr>
<td>Face validity</td>
<td>–</td>
<td>3.73 (0.76)</td>
<td>3.74 (0.77)</td>
<td>3.64 (0.77)</td>
<td>3.53 (0.73)</td>
</tr>
<tr>
<td>Procedural justice</td>
<td>–</td>
<td>3.74 (0.74)</td>
<td>3.66 (0.74)</td>
<td>3.66 (0.74)</td>
<td>3.71 (0.71)</td>
</tr>
</tbody>
</table>

Note: Honest baseline provided for comparison purposes only. Honest base line n = 57, all other conditions n = 100.
4.2.2. Impact of negatively worded warnings

Controlling for pre-warning scores, post-warning scores were significantly lower for extraversion ($t(198) = -2.41, p < .05$), conscientiousness ($t(198) = -2.88, p < .05$), and openness ($t(198) = -2.58, p < .05$) in the negative warning condition than scores in the no-warning condition. Post-warning scores for agreeableness ($t(198) = -2.71, p < .01$) and emotional stability ($t(198) = -2.76, p < .01$) were not significantly different from the no-warnings condition. This pattern of results provides partial support for Hypothesis 1 regarding negative warnings.

Although no specific hypotheses were specified, we also tested whether negative warnings were more effective than positive warnings. None of the differences between positive and negatively worded warnings were statistically significant.

4.2.3. Impact of accusations

Controlling for pre-warning scores, post-accusation scores were significantly lower than no-warning scores for extraversion ($t(198) = 4.20, p < .05$), agreeableness ($t(198) = 3.89, p < .05$), conscientiousness ($t(198) = 5.10, p < .05$), emotional stability ($t(198) = 3.61, p < .05$), and openness ($t(198) = 4.24, p < .05$). This pattern of results provides full support for Hypothesis 1 regarding accusations.

Hypothesis 2 stated that accusations would result in lower scores than those produced by either the positive or the negatively worded warnings. After controlling for pre-warning scores, post-accusation scores were significantly lower than post-positively worded warning scores for extraversion ($t(198) = 2.44, p < .05$), agreeableness ($t(198) = 3.49, p < .05$), and conscientiousness ($t(198) = 5.10, p < .05$), but not for emotional stability ($t(198) = 1.90, p > .01$) and openness ($t(198) = 1.78, p < .05$). After controlling for pre-warning scores, post-accusation scores were significantly lower than post-negative warning scores for agreeableness ($t(198) = 3.18, p < .05$) and emotional stability ($t(198) = 2.85, p < .05$), but not for extraversion ($t(198) = 1.76, p > .01$), conscientiousness ($t(198) = 2.22, p > .01$), or openness ($t(198) = 1.66, p > .01$). These results partially support Hypothesis 2 for the efficacy of accusations over positive and negative warnings.

4.3. Effect of warnings on test scores across levels of faking

Hypothesis 6 stated that the impact of warnings would depend on the level of faking that test-takers were engaging in, with warnings being more effective with test-takers engaging in a high level of distortion. We operationalized the level of faking two different ways. First, we used the number of bogus items that the test taker endorsed as a proxy of faking behavior. Second, we used the number of extreme responses endorsed on the impression management scale. Although the two measures are quite different, both are based on the assumption that more extreme levels of faking will result in higher scores (Burns & Christiansen, 2011). Across applicant instructions conditions 16.8% of the test-takers did not endorse any of the bogus items, 26.3% endorsed one, 21.5% endorsed two, 15.5% endorsed three, 8.0% endorsed four, and 12% endorsed all five bogus items; 18.0% of the test-takers endorsed no extreme responses on the impression management scale, 30.3% endorsed one extreme response, 25.5% endorsed two, 13.0% endorsed three, 6.5% endorsed four, and 6.8% endorsed all five.

To test Hypothesis 6 we altered our ANCOVA models to first account for pre-warning scores, then for differences across conditions, then for the measure of faking, and finally for the warnings condition and faking measure interaction. Operationalizing faking as responses to the bogus item scale, the number of bogus items endorsed was a significant covariate of post-warning scores for extraversion ($R(1,391) = 19.04, p < .05, \eta^2 = .05$) and openness ($R(1,391) = 14.62, p < .05, \eta^2 = .04$). This result indicates that participants endorsing more bogus items also had higher trait scores than those not endorsing these items. The number of bogus items endorsed did not moderate the effects of warning condition.

Using responses to the impression management scale as the measure of faking, the number of impression management items flagged was a significant covariate of post-warning scores for extraversion ($R(1,391) = 6.31, p < .05, \eta^2 = .02$), agreeableness ($R(1,391) = 7.24, p < .05, \eta^2 = .02$), conscientiousness ($R(1,391) = 9.27, p < .05, \eta^2 = .02$), and openness ($R(1,391) = 15.39, p < .05, \eta^2 = .04$). Again, these results indicate that participants endorsing extreme responses on the impression management scale had higher traits scores than those not endorsing these items. In addition, the number of impression management items flagged was a significant moderator of the effect of warning type for conscientiousness ($R(3,391) = 4.23, p < .05, \eta^2 = .03$) and openness ($R(3,391) = 3.87, p < .05, \eta^2 = .03$). As seen in Figs. 2 and 3, the impact of warnings on test scores was more pronounced as the level of pre-warning faking increased. These results partially support Hypothesis 6.

4.4. Effects of warnings on test-takers reactions

Based on the literature reviewed above we developed three specific hypotheses regarding warnings and test-taker reactions. Hypothesis 3 stated that negative warnings and accusations would
lead to increased anxiety; Hypothesis 4 stated that accusations would lead to decreased perceptions of fairness and accuracy; Hypothesis 5 stated that positive warnings would increase test-takers’ motivation. Unadjusted means for test-takers’ reactions are provided in Table 1.

Test-takers in the no-warning condition had lower levels of anxiety (M = 2.17, SD = .64) than participants in either the negative warnings (M = 2.34, SD = .79, t(198) = −1.66, d = −.23) and accusation conditions (M = 2.33, SD = .61, t(198) = −1.87, d = −.26), supporting Hypothesis 3. Accusations did not decrease perceptions of fairness but did influence perceptions of accuracy. Test-takers in the no-warnings condition thought the test was more accurate (M = 3.73, SD = .76) than test-takers in the accusation condition (M = 3.53, SD = .73, t(198) = 1.86, p < .05, d = −.26), partially supporting Hypothesis 4. Finally, participants in the positive warning condition had higher levels of motivation (M = 4.08, SD = .59) than participants in the no-warnings condition (M = 3.89, SD = .71, t(198) = −2.01, d = −.28), supporting Hypothesis 5.

To test Hypotheses 7 and 8, the impact of warnings on test-taker reactions would depend upon the level of faking, a set of ANCOVAs controlling for faking and the interaction between faking and warnings was conducted. Operationalizing faking as bogus items, the impact of warnings on perceptions of fairness depended upon the level of test-taker faking (F(3,392) = 3.20, p < .05, partial η² = .02). As shown in Fig. 4, perceptions of fairness were lower for test-takers engaging in more pre-warning distortion in the positive warnings and accusations conditions but the opposite pattern was observed for negative warnings. These results were partially consistent with Hypothesis 7, test-takers engaging in low levels of pre-warning faking will respond more positively to positive warnings than test-takers engaging in higher levels of faking. Inconsistent with Hypothesis 7, the opposite pattern was observed for negatively worded warnings. Perceptions of fairness were lower for test-takers engaging in lower levels of pre-warning distortion. The results for accusations were in the opposite direction than what was predicted for Hypothesis 8.

Changing the operationalization of faking to endorsement of impression management items, the ANCOVAs were rerun. The level of faking moderated the impact of warnings on test-taker perceptions for both perceived fairness (F(3,392) = 3.22, p < .05, partial η² = .02) and perceived accuracy (F(3,392) = 4.99, p < .05, partial η² = .04). As seen in Figs. 5 and 6, there was a positive association between level of faking and test-taker perceptions in the no-warnings condition but this trend was attenuated by introducing either positive or negative warnings and disappeared when accusations were used.

5. Discussion

Warnings have been proposed as a method of combating applicant distortion in contexts where test-takers might be motivated to distort their responses. Specifically, within applied samples warnings appear to keep some applicants from engaging in applicant distortion and retesting might be a solution for dealing with overly desirable response patterns. While research has provided some support for the use of warnings, the nature of manipulating test conditions has been a tricky proposition for practitioners meeting strict legal and ethical conditions of fairness. Additionally, research on accusations in applied personality assessment (Butcher et al., 1997; Cigrang & Staal, 2001; Ellingson et al., 2012) indicates that this is a strategy implied in practice, but very little is known about how it impacts test takers.

Turning to a laboratory setting, we manipulated the type of warning or accusation that participants received to examine the
effects of warnings on personality test scores. Our results indicate that warnings do produce noticeable changes in test-takers’ responses. Specifically, post-warning test scores tended to be lower in the warning conditions than in the no-warnings conditions. Consistent with Dwight and Donovan (2003), warnings containing consequences (negative warnings and accusations) tended to produce lower scores than warnings without consequences (positive warnings). These pre- to post-test differences in personality scores indicate that after receiving warnings, test takers began responding to the items in a less favorable manner. Our results also provided some additional information about the impact of warnings.

First, positive warnings resulted in the smallest change in scores and accusations resulted in the largest change. Note that there was considerable variation in the level of absolute change across personality traits. This is to be expected, as we would not expect all scores to decrease systematically – test takers do not fake all items systematically. Instead, test takers tend to create response patterns consistent with their perceived requirements of the position (Christiansen et al., 2005). These patterns of change are also consistent with results by Fan et al. (2012), who reported that warnings tended to decrease scores for perceived job relevant items more than perceived job irrelevant items. That is, positive warnings only reduced levels of openness compared to the no-warning group, with openness being the most highly distorted pre-warning trait in the current study. Negative warnings also reduced openness, extraversion, and conscientiousness as compared to the no-warning group, with extraversion being the second most highly distorted pre-warning trait as compared to the honest baseline.

Second, the impact of warnings might vary as a function of the level of pre-warning faking. Although participants who endorsed more bogus items and pre-warning impression management items tended to have higher levels of post-warning scores, the level of impression management moderated the effects of warnings for
two traits. Potentially reassuring, warnings had a larger impact on test takers illustrating higher levels of pre-warning faking than lower or average levels of distortion. However, even when pre-warning faking behaviors were low, warnings tended to reduce personality scores. These results are consistent with Fan et al. (2012) and Ellingson et al. (2012), who found that scores of even relatively honest test-takers tended to decrease after being warned or asked to retest. We found that these results were especially true for the accusation condition. This is concerning when considering targeted warnings, as current methods of detecting fakers are rife with false positives (Burns & Christiansen, 2006).

Our results regarding reactions to warnings were partially consistent with past results. Unique to warning studies, we found that positively worded warnings, although the least effective in reducing faking, tended to increase test-takers’ motivation. Consistent with Converse et al. (2008), we found that negative warnings increased test-takers’ levels of anxiety and that similar effects were found with accusations. Given that the stakes in our experiment (i.e., being selected to receive $50) are lower than the stakes that actual applicants face (i.e., being hired for a job), we would expect these results to be more robust in an applied sample.

Although past researchers have not reported any effects with warnings on perceived fairness or accuracy (McFarland, 2003), we found that accusations tended to reduce perceptions of test accuracy, but this depended in part on the level of distortion that test-takers were engaging in prior to warnings and the type of warning presented. Generally, participants engaging in more pre-warning faking tended to have higher perceptions of the simulation’s fairness and accuracy when they were not warned but described the test as less fair and accurate when they received a warning. Interestingly, perceptions of fairness and accuracy did not tend to change across warning conditions for test-takers engaging in low levels of pre-warning faking. In addition to providing information about the effects of warnings on applicant reactions, Figs. 4 and 5 suggest that absent warnings, test-takers engaging in higher levels of distortion will have more positive evaluations of the testing process. This indicates that test-takers engaging in faking behaviors actually view their test-taking strategy as a fair procedure and that they view warnings as removing this advantage.

Of particular concern to organizations should be the results regarding fairness perceptions in the negatively worded accusation condition (presented in Figs. 4 and 5). In this condition test takers engaging in low levels of faking perceived the process as less fair. Coupled with the increased anxiety caused by these types of warnings, we interpret these results as indicating that warnings might have detrimental effects on honest test-takers. Although past research has not identified fairness as a concern with warnings, we encourage future researchers to explore this area more and urge practitioners to carefully consider the impact of any warning procedures on all applicants.

Conclusions from the current study face certain limitations. Perhaps most importantly, while laboratory studies such as this provide opportunities to examine test-taker behavior in ways which would not be ethical in applied settings, the generalizability of the results is still questionable. Although students were certainly interested in receiving the $50 for being “hired,” this is a very different outcome than actually being hired by an organization. Similarly, students did not have to worry about the consequences of distorting their responses after the experimental session ended. Although we believe that such laboratory simulations provide a good opportunity to examine test-taker behavior, the above limitations might be especially relevant to examining test-taker reactions. We find it difficult to believe that computer delivered accusations of faking did not result in stronger differences in applicant reactions. This is possibly a result of some participants engaging in self-deception rather than intentional faking. Similar to participants engaging in impression management, participants engaging in self-deception will also endorse items that are unrealistically positive. However, self-deceiving participants are convinced that they are responding to these items in a truthful manner (Paulhus, 2002). These self-deceiving participants could potentially affect patterns of test-taker reactions. Unfortunately, in the current study we are not able to parse the data in such a way as to examine participants who are engaging in intentional faking versus self-deception.

Future research should continue to study the use of warnings in both applied and laboratory settings. While it is difficult to manipulate warnings in applied settings, they can be used consistently for cohorts of applicants. Such results can be seen with Vasilopoulos et al. (2005), where warnings were used at some hiring sites and not others. Information about reactions and test scores can be obtained and disseminated to provide information about differential effects of faking across different traits. Finally, although we report that accusations impact test takers engaging in high levels of faking the most, we are not endorsing this type of warning or suggesting that organizations implement accusations. Our results indicate that warnings impact the test scores of both those responding honestly and those engaging in faking and also increase the anxiety of test-takers. Although such accusations might be warranted in high stakes testing environments, we view current deficiencies in measures of detecting faking and the resulting problems in targeting testing takers as making such procedures ethically problematic.

References