Testing the Efficacy of a New Procedure for Reducing Faking on Personality Tests Within Selection Contexts

Jinyan Fan, Dingguo Gao, Sarah A. Carroll, Felix J. Lopez, T. Siva Tian, and Hui Meng

Online First Publication, January 16, 2012. doi: 10.1037/a0026655

CITATION
Testing the Efficacy of a New Procedure for Reducing Faking on Personality Tests Within Selection Contexts

Jinyan Fan
Hofstra University and Auburn University

Sarah A. Carroll
Professional Examination Service

T. Siva Tian
University of Houston

Dingguo Gao
Sun Yat-Sen University

Felix J. Lopez
Booz Allen Hamilton, Inc.

Hui Meng
East China Normal University

The authors propose a new procedure for reducing faking on personality tests within selection contexts. This computer-based procedure attempts to identify and warn potential fakers early on during the testing process and then give them a chance for recourse. Two field studies were conducted to test the efficacy of the proposed procedure. Study 1 participants were 157 applicants competing for 10 staff positions at a large university located in a southern city in the People’s Republic of China. In Study 1, potential fakers received a warning message, whereas nonfakers received a nonwarning (control) message. Study 2 participants were 386 Chinese college students applying for membership of a popular student organization at the same university where Study 1 was conducted. In Study 2, the warning and control messages were randomly assigned to all applicants. Results showed some promise for the proposed procedure, but several practical issues need to be considered.

Keywords: personnel selection, personality test, warnings, faking, online testing

During the last two decades, personality measures have been widely used as an employee-selection instrument, largely due to meta-analytic evidence that personality scores are predictive of important organizational outcomes (e.g., Barrick & Mount, 1991; Judge & Bono, 2001). Furthermore, theorists have argued (e.g., Hogan, Hogan, & Roberts, 1996), and empirical research has illustrated (e.g., Foldes, Duehr, & Ones, 2008), that personality measures typically do not display group mean score differences large enough to result in adverse impact.

However, there has been a consistent concern among selection professionals that self-report measures such as personality inventories are vulnerable to response distortion or faking (Rosse, Stecher, Miller, & Levin, 1998; White, Young, Hunter, & Rumsey, 2008). Although lab studies have demonstrated the detrimental effects of directed faking instructions on the validities of personality scores, field studies have yielded mixed results. Faking research has been an area of much debate (see Griffith & Peterson, 2006), and at present there is no indisputable evidence that faking is not a concern within actual selection contexts. We believe that faking should be examined further if industrial-organizational (I-O) psychologists are to advocate wider use of personality and other noncognitive measures in employment settings.

In this article, we introduce a new faking-mitigation procedure that blends the strengths of several different approaches that have been developed to address faking. We first define faking conceptually, then review major existing faking-mitigation strategies, and then present our new procedure. We specify several important evaluation criteria for the proposed procedure and then present two empirical studies, conducted with actual applicants in China, that tested its efficacy.

Conceptual Definition of Faking

One problem that has plagued faking research is the proliferation of terms associated with faking (e.g., "faking, response distor-
tion, impression management, socially desirable responding, dishonest responding, elevation, and several other terms), with few attempts to consolidate conceptual definitions of the faking construct. To start, one area of common ground among researchers is that they are interested in a general phenomenon of a job applicant providing responses to self-report measures that are more positive than his or her “true” or honest response. Although malingering is a form of faking that is relevant in clinical settings, personnel selection professionals are almost exclusively interested in “faking good.”

Paulhus (1984, 1991) proposed that socially desirable responding comprises at least two distinct dimensions: impression management (IM) and self-deceptive enhancement (SDE). Whereas IM refers to intentional response distortion, SDE refers to an unconscious response bias based on an overly positive self-image (Paulhus, 1984). It seems reasonable to suggest that most, if not all, personnel selection scholars and practitioners are interested in IM rather than SDE. Therefore, we define faking as the tendency to deliberately present oneself in a more positive manner than is accurate in order to meet the perceived demands of the testing situation. In our view, faking is intentional and situationally induced, and is therefore changeable.

Faking-Mitigation Strategies

Various faking-mitigation strategies in the literature can generally be classified into one of three paradigms. The first paradigm is reactive in that it allows faking to occur during the testing process and addresses it post hoc. Examples include statistical correction (e.g., Hough, Eaton, Dunnette, Kamp, & McCloy, 1990) and statistical modeling (e.g., Kuncel & Borneman, 2007; Smith & Ellingson, 2002; Zickar & Robie, 1999). The second paradigm is preventative and aims to prevent faking from occurring. Examples include subtle items (e.g., Holden & Jackson, 1981), forced-choice format (e.g., Jackson, Wrobleswki & Ashton, 2000), randomized item placement (e.g., McFarland, Ryan, & Ellis, 2002), speeded item presentation (e.g., Holden, Wood, & Tomaszewski, 2001), and warnings against faking (e.g., Dwight & Donovan, 2003). Research evidence has been mixed regarding the effectiveness of these strategies (for a review, see Kuncel & Borneman, 2007).

The third paradigm is an emerging one that combines reactive and preventative strategies and centers on what we call the “test-warning-retest” procedure. The original idea can be traced back to several personality test manuals (e.g., 16 personality factor [16PF] questionnaire) that have recommended retesting individuals who have suspicious protocols that have likely been faked. This paradigm has remained largely neglected by scholars until recently (e.g., Butcher, Morfitt, Rouse, & Holden, 1997; Ellingson & Heggestad, 2004; Landers, Sackett, & Tuzinski, 2011). For example, Butler et al. administered a paper-and-pencil version of the Minnesota Multiphasic Personality Inventory–2 (MMPI-2) to 271 airline pilot applicants. Seventy-two MMPI-2 profiles were ruled invalid, due to extreme scores on the L and/or K scales suggesting underreporting psychopathology. Butler et al. gave these 72 applicants a polite warning and asked them to complete the MMPI-2 again. Results showed that 57 of the 72 applicants returned valid, interpretable profiles the second time.

Landers et al. (2011) implemented a “test-warning-retest” procedure in an online personality test context. In their study, applicants who displayed blatant extreme responding (BER; using only extreme responses, i.e., all 1s and 5s on a Likert scale) early in the testing process were identified as potential fakers, and were sent a computerized warning in real time. The warned respondents had the option of either continuing the test or returning to earlier items and changing their responses. Landers et al. reported that their procedure lowered the occurrence of BER in a large sample of applicants for management positions at a nationwide retailer.

In the present research, we propose a different “test-warning-retest” procedure, which blends strengths of previous similar procedures, for instance, Butcher et al.’s (1997) polite warning principle, Landers et al.’s real-time warning idea, and Ellingson and Heggestad’s (2004) research design. The proposed procedure is described below.

The Proposed Procedure

Like Landers et al. (2011), the proposed procedure is implemented in a computer-based testing system so that test-takers’ scores can be calculated and communicated to them in real time. As can be seen in Figure 1, the proposed procedure consists of two major steps. In the first step, a test-taker completes items in the initial block via a computer. The initial block contains these measures: (a) a bogus statement (BS) scale, (b) Paulhus’ (1998) IM scale, and (c) a small sample of personality items. The first two measures are used to identify potential fakers.

Bogus statements attempt to catch a test-taker engaging in deliberate deception (e.g., Anderson, Warner, & Spector, 1984; Dwight & Donovan, 2003). If a test-taker feigns knowledge of, or experience with, too many nonexistent concepts or techniques described in the bogus statements, he or she is considered to be faking. The IM scale tries to identify more subtle fakers. IM items tap into relatively uncommon good behaviors and relatively common bad behaviors. If a test-taker’s IM score exceeds a cut-off score that has been deemed, based on normative data, to be highly unlikely to be obtained under normal circumstances, he or she is considered to be engaging in faking. The legitimacy of this rationale is greatly strengthened in real selection contexts where disproportionate applicants are expected to have IM scores that exceed the cut-off score (e.g., Fan & Wong, 2009; Rosse et al., 1998).

If a test-taker’s score in either the BS or the IM scale exceeds certain preset criterion, he or she will be flagged as a potential faker and will receive a warning message before being brought to the second step, the main block. The main block contains the same BS and IM scales as the initial block, and the full set of personality items. Conversely, if a test-taker’s BS and IM scores do not exceed the preset criterion, he or she will be classified as a nonfaker, will receive a nonwarning (control) message, and will then complete the same main block items as potential fakers. All items in the initial block are retested in the main block for all test-takers.

We note that use of warnings is not a new idea, as decades of research have shown that warnings may reduce faking behaviors (Dwight & Donovan, 2003). The uniqueness of the proposed procedure lies in the automated process and the opportunity for retesting.

Evaluation Criteria

We specify three major criteria to evaluate the efficacy of the proposed procedure. First, we expect a Treatment $\times$ Time inter-
action effect on scales included in the initial block. That is, the flagged and warned group should lower their scores (i.e., having less favorable scores) from Time 1 (the initial block) to Time 2 (the main block), whereas the nonflagged and unwarmed group should have lower (less favorable) scores than the flagged and warned group at Time 1, but should maintain similar scores over time (Hypothesis 1a). However, one qualification is that the above interaction effect should be much stronger for the IM and BS scales, and for job-relevant personality scales than for personality scales with low job relevance (Hypothesis 1b). We anticipate this because we expect applicants to fake less on job-irrelevant scales than on job-relevant scales (Rosse et al., 1998; Vasilopoulos, Cucina, & McElreath, 2005). Consequently, flagged and warned applicants should have smaller score reductions on job-irrelevant personality scales; at the same time, nonflagged and unwarmed applicants should have similar scores over time on the same job-irrelevant scales, yielding weaker Treatment × Time interactions. Note that because potential fakers are identified based on IM or BS scores, regression toward the mean might also predict a similar Treatment × Time interaction on IM and BS scales; however, such an effect cannot explain the hypothesized interaction effects on job-relevant personality scales.

Second, one crucial feature of the proposed procedure is efficiency through the use of the initial block. That is, we believe there is no need to place the entire set of personality items in the initial block; a small sample of personality items will achieve the desired effect. To establish efficiency, we must show that the above warning effect carries over to those scales not included in the initial block. That is, we expect the mean score differences between the two treatment groups on nonretested personality scales (particularly the job-relevant ones) to be comparable to the mean score differences on retested scales at Time 2 (Hypothesis 2).

Third, selection professionals often face a difficult scenario in which potential fakers are overly represented among the top applicants (Goffin & Christiansen, 2003). The proposed procedure, which targets potential fakers and warns only them, should help address this problem. To test this effect, we first rank all test-takers on the basis of their personality scores at Time 1 and calculate the percentage of potential fakers at various selection ratios. Next, we rank all test-takers and calculate the percentages of fakers again, this time using the same set of personality scores obtained at Time 2. We expect a lower percentage of potential fakers in the hired group when Time 2 scores are used than when Time 1 scores are used (Hypothesis 3a). Furthermore, based on prior research (e.g., Mueller-Hanson, Heggestad, & Thornton, 2003; Rosse et al., 1998), we expect the above benefit to be more evident as selection ratio decreases (Hypothesis 3b).

Effects on Applicant Perceptions

We also investigate how the proposed procedure would influence applicant perceptions, specifically, three fairness-related perceptions (perceived test fairness, perceived face validity, and test satisfaction) and one attitudinal perception (test-taking motivation). Receiving a warning is an unfavorable outcome, which has been shown to lead to negative applicant perceptions (Ryan & Ployhart, 2000). However, we argue that the proposed procedure might mitigate these negative perceptions. First, the warning is not a final outcome; flagged and warned applicants are given an opportunity for recourse. This reconsideration opportunity should boost fairness perceptions (Gilliland, 1993). Second, the warning message was carefully developed on the basis of interactional justice principles such as adequate explanations and interpersonal civility (Bies & Moag, 1986). Considerable research has shown that high interactional justice buffers the negative effects of unfavorable outcomes (e.g., Greenberg, 1990). Thus, we hypothesize that although flagged and warned applicants should have less positive perceptions than nonflagged and unwarmed applicants, the differences should not be statistically significant (Hypothesis 4).
Study 1

Method

Sample and procedure. Participants were 157 applicants competing for 10 staff positions at a large university in a southern city in the People’s Republic of China. In this sample, 49 (31%) were male, the average age was 26 years old, and all had at least a master’s degree. Participants’ specialty backgrounds were very diverse. These applicants were invited to participate in an onsite psychological assessment that took place at a classroom equipped with computer facilities. As can be seen in Figure 1, once logged in, applicants first saw a welcome message, followed by a preview of the testing procedure. They then completed a brief demographic survey. Next, they proceeded to the initial block, which consisted of a BS scale (framed as a school activity survey), an IM scale, and three Chinese 16PF scales (see the Measures section for rationales of selecting these scales). Applicants whose scores on the BS or IM scale exceeded the preset faking criterion \(n = 45\) received the following warning message (translated from Chinese):

Thank you for participating in this portion of the selection process. However, we have noticed some unusual response patterns in your answers and wish to clarify the issue. The personality inventory and the school activity survey which you are completing have two embedded social desirability scales. These scales identify people who might have tailored their responses to what they believe the hiring organization wants to hear, in order to increase the chances of getting the job.

Your response profile up to this point is similar to that of someone who is known to be answering in a socially desirable way. We do not intend to insult your integrity; we only want to get a clear understanding of who you are. Inaccurate information from the assessment, if used as the basis for selection, may result in poor person-job fit and/or poor person-organization fit. This may further lead to unfit employees and a low employee satisfaction, feeling inadequate, dissatisfied, having decreased motivation and eventually quitting; or being terminated by the organization. Thus, we would like to underscore the importance of total honesty in completing these inventories.

That said, we would like to offer you an opportunity to complete the inventories all over again. Remember, be yourself and answer each question as it best describes you. Finally, rest assured that your previous responses on these inventories will NOT be considered in our final selection decisions. However, we have found in the past that some candidates had repeatedly distorted their response. These individuals were quickly discovered and were immediately removed from the selection process.

Conversely, applicants \(n = 112\) whose BS and IM scores did not exceed the preset faking criterion received the following control message (translated from Chinese):

Thank you for participating in this portion of the selection process. A random system check indicates the testing system is working well. Please continue the test. Be reminded that as part of the testing procedure, some of the items will be presented twice. So don’t be surprised if you see some of the items showing up again on the screen.

Both groups then proceeded to the main block, which contained the same BS and IM items and the full set of Chinese 16PF items. After finishing the main block items, applicants completed several perception measures before seeing the “Thank You” page, which concluded the assessment. Applicants were told that their answers to the perception measures would not be used for selection decisions. Applicants’ scores on several dimensions in the main block were calculated, and 20 applicants were invited for an onsite interview, which determined who was offered the job. The hiring university decided to eliminate three applicants whose IM and BS scores exceeded the preset faking criteria at both the initial block and the main block.

Measures.

Personality test The 185-item Chinese 16PF (Dai & Zhu, 1988) was used, which is a translated and an adapted version of the fourth edition 16PF (Cattell, Eber, & Tatsuoka, 1970). The Chinese 16PF is one of the most widely used personality inventories in both research and applied settings in China. According to its manual (Dai & Zhu, 1988) and several recent large-scale validation studies (e.g., Song, Han, & Liu, 2006; Yang et al., 2007), Chinese 16PF scores exhibits good psychometric properties. For instance, alpha coefficients for the 16 dimension scales ranged from mid .60s to low .80s, and 2-week test–retest coefficients ranged from low .60s to low .80s.

Relevance and direction rating. Six Chinese I-O graduate students, who had extensive training in using the Chinese 16PF, served as subject matter experts (SMEs). These SMEs, with no knowledge of the proposed procedure, rated both the relevance and favorable direction of Chinese 16PF dimensions to the university staff position. A 5-point scale for relevance ratings was used, ranging from 1 (irrelevant) to 5 (extremely relevant), and a 3-point scale for direction ratings was used such that “+” means higher scores (i.e., higher amount of the trait) are more favorable than lower scores, “−” means lower scores are more favorable, and “?” means the favorable direction is unclear. Table 1 lists the averaged relevance rating and aggregated favorable direction rating for each 16PF dimension. (Factor B Reasoning was excluded, because it is a rough measure of cognitive ability, not personality.) In aggregating direction ratings, the rule that a dimension’s favorable direction was considered clear only if at least five out of six SMEs had the same rating was adopted. On the basis of these ratings, the choice was made to place the Factor C Emotional Stability scale, the Factor G Rule-Consciousness scale, and the Factor N Private-ness scale in the initial block. Factors C and G represent traits that applicants most likely want to fake on (very high job relevance and a clear favorable direction), whereas Factor N represents traits that applicants least likely want to fake on (relatively low job relevance and an unclear favorable direction).

Bogus statements. Dwight and Donovan’s (2003) BS scale was adapted for the present study. Dwight and Donovan used two bogus statements, one of which was feared might not suit the Chinese culture. So, a new bogus statement was written to replace the problematic one. Thirteen genuine items were also written. These 15 items formed the BS scale. Respondents were asked to indicate how often they had engaged in these behaviors while at college. Items were rated on a 4-point scale ranging from 1 (never) to 4 (often). For the two bogus statements, responses endorsing 1 (never) were assigned 0 points, and all other responses were assigned 1 point.

In a pilot study, two classes of undergraduate psychology juniors at an East Coast university in China completed the BS scale separately, under two different conditions. In the honest condition \(n = 56\),
students were asked to respond honestly. In the motivated condition
(n = 52), Griffith, Chmielowski, and Yoshita’s (2007) deception
procedure was followed. The experimenter presented herself as a
recruiter working for a local consulting firm who was recruiting
several summer interns. She asked students to complete a screening
questionnaire (the BS scale) and then debriefed them. All participants
received 5 Chinese Yuan (approximately $0.78 U.S.) for their time
(10 min). Analyses indicated that (a) the motivated group had a higher
mean BS score than the honest group (0.55 vs. 0.09), t(106) = 4.22,
p < .01, d = 0.90, (b) whereas in the honest condition, nobody
endorsed two BS items; in the motivated condition, eight (15%) endor-
sed both BS items; and (c) in the honest condition, the two-item
BS scale had a coefficient alpha of .07, whereas in the motivated
condition the BS scales had a coefficient alpha of .60. Thus, the BS
scale seemed to function well. A BS score of 2 (the maximum
possible score) was used as the cut score for determining who
would be flagged as potential fakers.

Impression management. The Chinese version (Fan, Wong,
Carroll, & Lopez, 2008) of the IM scale from the Balanced
Inventory of Desirable Responding (Paulhus, 1998) was used to
measure intentional response distortion. Fan et al. translated and
adapted English items into Chinese and validated the IM scale in
Chinese populations. To establish the norm of Chinese IM scores
for the present research, Fan et al.’s (2008) data were used.
Participants in Fan et al.’s study were 4,645 first-year under-
graduate and graduate students in a Chinese university who went
through a mandatory psychological assessment (which included
the Chinese 16PF and the Chinese IM scale). Because the assess-
ment took place after these students had entered into this univer-
sity, they presumably had little motivation to fake. In this norma-
tive sample, Chinese IM scores had a mean of 8.79 and a standard
deviation of 4.10. Thus, in the present research applicants endors-
ing 17 or more out of 20 IM items (equivalent to 2 SDs above
the normative mean) were flagged as potential fakers. The coefficient
alpha was .86 at both time points (i.e., initial block and main
block) in Study 1.

Perceived test fairness. Perceived test fairness was measured
with eight items taken from Tonidandel, Quinones, and Adams
(2002) and two items from Smither, Reilly, Millsap, Pearlman,
and Stoffey (1993). Back-translation (cf. Brislin, 1970) was used to
translate the original English items into Chinese. Items were rated
on a 7-point Likert scale ranging from 1 (strongly disagree) to 7
(strongly agree). One sample item is “Overall, I believe the test
was fair.” The coefficient alpha was .83 in this sample.

Table 1
Winsorized Means, Winsorized (Standard Deviations), Standardized Differences, and SME Ratings in Study 1

<table>
<thead>
<tr>
<th>Measure</th>
<th>Flagged and warned (n = 45)</th>
<th>Nonflagged and unwarned (n = 112)</th>
<th>SME ratings</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Time 1</td>
<td>Time 2</td>
<td>d_{ws2}</td>
</tr>
<tr>
<td>Social desirability scales</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Impression Management</td>
<td>17.59 (2.73)</td>
<td>10.63 (2.80)</td>
<td>1.52&quot;</td>
</tr>
<tr>
<td>Bogus Statement</td>
<td>1.30 (0.89)</td>
<td>0.19 (0.47)</td>
<td>.83&quot;</td>
</tr>
<tr>
<td>Chinese 16PF scales</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Factor A Warmth</td>
<td>14.37 (2.76)</td>
<td>12.66 (2.41)</td>
<td>.68*</td>
</tr>
<tr>
<td>Factor C Emotional Stability</td>
<td>20.96 (2.32)</td>
<td>18.57 (2.01)</td>
<td>1.39&quot;</td>
</tr>
<tr>
<td>Factor E Dominance</td>
<td>12.30 (2.47)</td>
<td>13.66 (1.93)</td>
<td>.17</td>
</tr>
<tr>
<td>Factor F Liveliness</td>
<td>16.48 (3.12)</td>
<td>17.47 (2.64)</td>
<td>.65*</td>
</tr>
<tr>
<td>Factor G Rule-Consciousness</td>
<td>15.96 (1.60)</td>
<td>14.10 (1.88)</td>
<td>1.65&quot;</td>
</tr>
<tr>
<td>Factor H Social Boldness</td>
<td>14.11 (3.64)</td>
<td>14.63 (2.57)</td>
<td>.18</td>
</tr>
<tr>
<td>Factor I Sensitivity</td>
<td>12.48 (2.70)</td>
<td>11.84 (2.31)</td>
<td>.26</td>
</tr>
<tr>
<td>Factor L Vigilance</td>
<td>7.74 (1.94)</td>
<td>7.06 (1.59)</td>
<td>.21</td>
</tr>
<tr>
<td>Factor M Abstractedness</td>
<td>13.44 (1.59)</td>
<td>13.50 (1.92)</td>
<td>.03</td>
</tr>
<tr>
<td>Factor N Privateness</td>
<td>8.56 (1.24)</td>
<td>9.12 (1.21)</td>
<td>-.84&quot;</td>
</tr>
<tr>
<td>Factor O Apprehension to Change</td>
<td>8.30 (2.67)</td>
<td>6.09 (1.90)</td>
<td>.90**</td>
</tr>
<tr>
<td>Factor Q, Emotional Openness to Change</td>
<td>10.56 (1.58)</td>
<td>11.09 (1.53)</td>
<td>.34</td>
</tr>
<tr>
<td>Factor Qz, Self-Reliability</td>
<td>8.93 (1.59)</td>
<td>9.75 (1.57)</td>
<td>.52</td>
</tr>
<tr>
<td>Factor Qz, Perfectionism</td>
<td>14.03 (1.67)</td>
<td>14.29 (1.48)</td>
<td>.17</td>
</tr>
<tr>
<td>Factor Q, Tension</td>
<td>7.89 (2.09)</td>
<td>7.35 (2.00)</td>
<td>.27</td>
</tr>
<tr>
<td>Applicant Perceptions Scales</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perceived Test Fairness</td>
<td>4.68 (0.36)</td>
<td>4.72 (0.39)</td>
<td>.10</td>
</tr>
<tr>
<td>Perceived Face Validity</td>
<td>4.96 (0.48)</td>
<td>5.30 (0.53)</td>
<td>.66&quot;</td>
</tr>
<tr>
<td>Test Satisfaction</td>
<td>4.37 (0.47)</td>
<td>4.44 (0.43)</td>
<td>.16</td>
</tr>
<tr>
<td>Test-Taking Motivation</td>
<td>4.37 (0.44)</td>
<td>4.47 (0.39)</td>
<td>.25</td>
</tr>
</tbody>
</table>

Note. The degree of winsorizing was 20%. d_{ws2} = within-subject d in the warning condition; d_{ws2} = within-subject d in the control condition. For d_{ws2} and d_{wb2}, positive values indicate mean scores were higher in Time 1 than mean scores in Time 2. d_{wb2} = between-subject d (the warning condition vs. the control condition) at Time 2. For d_{wb2}, positive values indicate the faked and warned group had higher mean scores than the nonfaked and unwarned group at Time 2. SME = subject matter expert; R = relevance; D = direction; “+” means higher amount of the trait is favorable; “−” lower amount of the trait is favorable; “?” means direction is unclear.

*p < .05. **p < .01.
Perceived face validity. The five-item questionnaire developed by Smither et al. (1993) was used to measure perceived face validity. The original items were translated into Chinese using the back-translation technique. Items were rated on a 7-point Likert scale ranging from 1 (strongly disagree) to 7 (strongly agree). One sample item is “I did not understand what the test has to do with the job.” The coefficient alpha was .86 in the present sample.

Test satisfaction. The eight-item inventory developed by Tonidandel et al. (2002) was used to measure applicants’ satisfaction with the testing process. The original items were translated into Chinese using the back-translation technique. Items were rated on a 7-point Likert scale ranging from 1 (strongly disagree) to 7 (strongly agree). One sample item is “So far, participation in the testing process has been a positive experience.” The coefficient alpha was .88 in this sample.

Test-taking motivation. The 10-item Motivation subscale of the Test Attitude Scale by Arvey, Strickland, Drauden, and Martin (1990) was used to measure applicants’ test-taking motivation. The original items were translated into Chinese using the back-translation technique. Items were rated on a 7-point Likert scale ranging from 1 (strongly disagree) to 7 (strongly agree). One sample item is “I wanted to do well on this test.” The coefficient alpha was .77.

Results

Treatment × Time interactions. Due to concerns with covariance heterogeneity, uneven sample sizes (n = 45 vs. n = 112) and different skewness across the two treatment groups, we used trimmed-mean analysis of variance (ANOVA), a form of robust ANOVA (Wilcox, 2005), to test the hypotheses. Robust ANOVAs were conducted in the free software R using the functions written by Wilcox. We also conducted Yuen’s t tests (a robust t test) in each treatment condition separately. Table 1 presents 20% winsorized means and standard deviations of all scales at Time 1 and Time 2 based on the winsorized sample, standardized mean differences over time (within-subject ds), and across different treatment conditions (between-subject ds) as appropriate.

Hypothesis 1 predicted (a) a Treatment × Time interaction effect on scales in the initial block and (b) a weaker interaction effect on the Factor N Privateness scale than on the other four scales. For the IM scale, the Treatment × Time interaction effect was significant, F(1, 72) = 29.24, p < .01, partial η² = .29. Yuen’s t tests indicated that the flagged and warned group significantly lowered the IM mean score over time (17.59 vs. 11.93), t(26) = 5.33, p < .01, d = 1.52, whereas the mean IM scores for the nonflagged and unwarned group changed little over time (10.63 vs. 10.82), t(67) = -0.91, p = .37, d = -0.15. Figure 2 shows that the flagged and warned group had a higher IM mean score than the nonflagged and unwarned group at Time 1 (initial block), but the difference became much smaller at Time 2 (main block). Thus, Hypothesis 1a was supported for the IM scale. For the BS scale, the interaction was significant, F(1, 81) = 5.80, p < .05, partial η² = .07. The flagged and warned group decreased the BS mean score over time (1.30 vs. 0.85), t(26) = 3.29, p < .01, d = 0.83, whereas the non flagged and unwarned group showed a smaller score reduction (.19 vs. .09), t(67) = 2.37, p < .05, d = 0.36. Thus, Hypothesis 1a received partial support for the BS scale.

Figure 2. The Time × Warning interaction effect on the impression management scale in Study 1. The y-axis runs approximately between one standard deviation below and above the overall mean.

For the Factor C Emotional Stability scale, the interaction effect was significant, F(1, 76) = 30.94, p < .01, partial η² = .29. For Factor G Rule-Consciousness scale, the interaction effect was also significant, F(1, 100) = 38.14, p < .01, partial η² = .28. According to Table 1, for Factors C and G scales, the flagged and warned group lowered the mean score over time (d = 1.39 and 1.65, respectively; ps < .01), whereas the nonflagged and unwarned group had little mean score change over time (d = -0.20 and -0.20; p = .28 and .22, respectively). The Treatment × Time interaction formats of the Factors C and G scales were very similar to that of the IM scale. Thus, Hypothesis 1a was supported for both the Factor C scale and the Factor G scale.

For the Factor N Privateness scale, the interaction was significant, F(1, 103) = 6.53, p < .05, partial η² = .06. The flagged and warned group increased their mean score over time (8.56 vs. 9.30), t(26) = -3.09, p < .01, d = -0.84, whereas the nonflagged and unwarned group had little score change over time (9.12 vs. 9.17), t(67) = -0.34, p = .74, d = -0.06. Comparisons of the interaction effects and the warning effects among the five initial block scales indicated that these two effects were both smaller for the Factor N Privateness scale than for the IM, Factor C, and Factor G scales. One exception was the BS scale, for which the two effects were comparable to that of the Factor N scale. Taken together, Hypothesis 1b received some support.

For comparison purposes, we repeated the above analyses using the regular two-way mixed design ANOVA. Results based on the regular ANOVA led to the same statistical conclusions for all scales, with one exception: For the Factor N scale, the Treatment × Group interaction was not significant, F(1, 155) = 2.07, p = .15, partial η² = .01. Thus, were the regular ANOVA used, Hypothesis 1b would have received stronger support.

The carry-over effect. Hypothesis 2 predicted that the warning effect should carry over to 16PF scales not included in the initial block. Table 1 indicates that at Time 2 (main block), the flagged and warned group had somewhat less favorable scores in Factor C Emotional Stability and Factor G Rule-Consciousness.
scales than the nonflagged and unwarned group; however, the group differences were nonsignificant. Thus, we expected a similar pattern of results on nonretested scales. Most important, the carryover effect should be observed on 16PF dimensions deemed relevant to the staff position (i.e., Factor A Warmth, Factor I Sensitivity, Factor L Vigilance, Factor O Apprehension, Factor Q3 Perfectionism, and Factor Q4 Tension).

Consistent with our prediction, the flagged and warned group tended to have similar or less favorable (yet nonsignificant) scores on the above 16PF scales than the nonfaked and unwarned group (see Table 1), but with two exceptions. First, according to Yuen’s $t$ test, the flagged and warned group continued to have a significantly more favorable mean score than the nonflagged and unwarned group on the Factor A Warmth scale ($14.37$ vs. $12.66$), $t(42) = 2.17$, $p < .05$, $d = 0.68$. Second, the former group had a significantly less favorable mean score on the Factor O Apprehension scale than the latter group ($8.30$ vs. $6.09$), $t(37) = 3.02$, $p < .01$, $d = 0.90$. Interestingly, carry-over effects were also observed for the remaining 16PF scales whose favorable directions were clear, that is, Factor F Liveliness, Factor H Social Boldness, and Factor M Abstractedness. Thus, Hypothesis 2 received some support.

**Effects on individual hiring decisions.** Hypothesis 3 predicted that the proposed procedure may help reduce overrepresentation of potential fakers among the top applicants. Table 2 lists the percentage of potential fakers in the hired group at various selection ratios in a top-down selection schema for two scenarios: Scenario A, when the sum of Time 1 standardized scores of the Factor C Emotional Stability and Factor G Rule-Consciousness scales are used to rank applicants, and Scenario B, when the sum of Time 2 standardized Factors C and G scores are used to rank applicants.

Several observations seem warranted from Table 2. First, consistent with prior research (e.g., Rosse et al., 1998), as the selection ratio decreased, the percentage of potential fakers in the hired group increased. This phenomenon occurred in both scenarios. Second, compared with Scenario A (the baseline), Scenario B (the proposed procedure) had lower percentages of potential fakers in the hired group across all selection ratios. Third, this benefit was most evident at the smallest selection ratios of 10% and 5%. Finally, the last column indicates that as selection ratios decreased, the percentage of applicants flagged at Time 1 as potential fakers who would be hired at Time 2 also decreased. This provides convergent evidence that the proposed procedure is somewhat effective at reducing the number of potential fakers among the top applicants, particularly at small selection ratios. These findings strongly supported Hypotheses 3a and 3b.

**Effects on applicant perceptions.** Table 1 shows that the flagged and warned applicants tended to report less positive perceptions of the testing experience than the nonflagged and unwarned applicants. However, only one of the four perceptions, perceived face validity, showed a significant group mean difference in Yuen’s $t$ test ($4.96$ vs. $5.30$), $t(51) = -2.36$, $p < .05$, $d = -0.66$. For the other three perceptions, group differences were not significant and the effect sizes were small (see Table 1). Therefore, Hypothesis 4 received some support.

**Discussion**

The results of Study 1 provided some initial evidence for the efficacy of the proposed procedure. Flagged applicants, after receiving the warning message, lowered their personality scores, whereas nonflagged applicants, after receiving the control message, had little score change. It is tempting, based on the above results, to attribute the score reduction effect to the warning message; however, caution needs to be exercised, as several alternative explanations such as regression toward the mean and the Maturation × Treatment interaction cannot be ruled out.

The finding that the Warning × Time interaction was weaker on the Factor N Privateness scale, a scale with low job relevance and an unclear favorable direction, supports the notion that job applicants do not display a similar amount of faking across all person-

---

**Table 2**

*The New Procedure’s Effects on Individual Hiring Decisions*

<table>
<thead>
<tr>
<th>SR</th>
<th>No. hired</th>
<th>If T1 Factors C &amp; G scores were used for selection</th>
<th>Potential fakers</th>
<th>If T2 Factors C &amp; G scores were used for selection</th>
<th>Potential fakers</th>
<th>% difference</th>
<th>% of T1 fakers hired at T2</th>
</tr>
</thead>
<tbody>
<tr>
<td>.05</td>
<td>9</td>
<td>$n$ 8</td>
<td>$%$ 89</td>
<td>$n$ 3</td>
<td>$%$ 33</td>
<td>56</td>
<td>38</td>
</tr>
<tr>
<td>.10</td>
<td>15</td>
<td>11</td>
<td>73</td>
<td>7</td>
<td>47</td>
<td>26</td>
<td>45</td>
</tr>
<tr>
<td>.20</td>
<td>30</td>
<td>16</td>
<td>53</td>
<td>11</td>
<td>35</td>
<td>18</td>
<td>50</td>
</tr>
<tr>
<td>.30</td>
<td>48</td>
<td>22</td>
<td>46</td>
<td>14</td>
<td>30</td>
<td>16</td>
<td>55</td>
</tr>
<tr>
<td>.40</td>
<td>63</td>
<td>29</td>
<td>46</td>
<td>16</td>
<td>25</td>
<td>21</td>
<td>45</td>
</tr>
<tr>
<td>.50</td>
<td>79</td>
<td>35</td>
<td>44</td>
<td>18</td>
<td>23</td>
<td>21</td>
<td>54</td>
</tr>
<tr>
<td>.60</td>
<td>94</td>
<td>36</td>
<td>38</td>
<td>20</td>
<td>21</td>
<td>17</td>
<td>67</td>
</tr>
<tr>
<td>.70</td>
<td>110</td>
<td>40</td>
<td>36</td>
<td>22</td>
<td>20</td>
<td>16</td>
<td>75</td>
</tr>
<tr>
<td>.80</td>
<td>126</td>
<td>42</td>
<td>33</td>
<td>25</td>
<td>20</td>
<td>13</td>
<td>86</td>
</tr>
<tr>
<td>.90</td>
<td>141</td>
<td>44</td>
<td>31</td>
<td>25</td>
<td>18</td>
<td>13</td>
<td>86</td>
</tr>
<tr>
<td>1.00</td>
<td>157</td>
<td>45</td>
<td>29</td>
<td>28</td>
<td>18</td>
<td>11</td>
<td>100</td>
</tr>
</tbody>
</table>

*Note.* SR = selection ratio; Factor C = Emotional Stability; Factor G = Rule-Consciousness; % difference = difference in the percentage of potential fakers in the hired group across the two scenarios; T1 = Time 1 (initial block); T2 = Time 2 (main block).
ality scales; rather, they fake more on job-relevant scales than on job-irrelevant ones (Rosse et al., 1998; Vasilopoulos et al., 2005). The reason for the relatively weak interaction effect on the BS scale was likely due to the small range of BS scores (0–2) and the low percentage of applicants endorsing bogus statements. In other words, we might have encountered a floor effect.

The finding that the warning effect carried over to many non-retested job-relevant 16PF scales suggested that placing a small sample of personality items in the initial block was sufficient in achieving the warning effect for all personality scales. This efficiency should be attractive to selection practitioners, who are aware that both applicants and hiring organizations prefer a shorter assessment time. The finding that the proposed procedure reduced the “fakers-rising-to-the-top” problem should also be appealing to practitioners, who have long struggled with this problem with few viable solutions. In addition, the proposed procedure had only small effects on several applicant perceptions, which should be encouraging to selection professionals who might consider implementing the proposed procedure in their practice.

Despite these positive findings, it is important to note that the proposed procedure helped reduce faking, but not eliminate it. For instance, the flagged and warned group continued to have a significantly higher mean score on the Factor A Warmth scale than the nonflagged and unwarned group in the main block. There were still around 30%–40% of potential fakers identified in the initial block, who would be hired at the main block even at the smallest selection ratios.

Finally, from a research design perspective, Study 1 was a quasi-experiment, not a true experiment, which requires incorporation of two additional conditions (cells): one in which flagged applicants receive the control message and the other in which nonflagged applicants receive the warning message. Although a true experiment is impractical (or even unethical) within real selection contexts, it would help shed light on the psychological mechanisms behind the proposed procedure (e.g., internal and construct validity). Therefore, we conducted a true field experiment in Study 2.

**Study 2**

The random assignment of the warning and control messages to all applicants yielded four cells in Study 2: (a) flagged and unwarned, (b) flagged and warned, (c) nonflagged and unwarned, and (d) nonflagged and warned. The primary purposes of Study 2 were threefold. The first purpose was to establish the internal validity of the proposed procedure. If Cell a applicants (flagged and unwarned) have little score change over time on scales included in the initial block, aforementioned alternative explanations such as regression toward the mean and Maturation × Group interaction can be ruled out.

The second purpose of Study 2 was to examine the construct validity of the proposed procedure. One aspect of construct validity requires that the warning should lead to applicants responding more honestly to personality items rather than simply responding to the experimental demand. Ideally, we would like to see that Cell d applicants (nonflagged and warned) do not lower their personality scores significantly. However, such an expectation is complicated by the fact that in field settings, faking criteria are typically set very high to avoid false positives. That is, there are mild fakers among nonflagged applicants who, after being warned, would respond more honestly by lowering their scores. If so, the amount of score reduction should be smaller among Cell d applicants (nonflagged and warned) than among Cell b applicants (flagged and warned); this is because the former group should fake less. Conversely, if Cell d applicants have a similar amount of score reduction as Cell b applicants, then this would strongly favor the experimental demand explanation, which would challenge construct validity.

On the basis of the above discussions, we hypothesize a three-way Treatment (warning vs. control) × Group (potential fakers vs. nonfakers) × Time interaction effect on scales included in the initial block (Hypothesis 1a). Applicants receiving the control message should not change their scores much over time, regardless of their group membership. Applicants receiving the warning message should lower their scores; however, the amount of score reduction should be smaller among nonflagged applicants than among flagged applicants. We further suggest that the three-way interaction should be weaker on the Factor N Privateness scale than on other initial block scales (Hypothesis 1b), based on the similar argument made in Study 1.

Like in Study 1, we hypothesize that the above Treatment × Group interaction effect should carry over to nonretested personality scales in the main block. That is, for nonretested personality scales, applicants receiving the warning message should have less favorable scores than applicants receiving the control message, but the warning effect should be smaller among nonflagged than among flagged applicants (Hypothesis 2a). Furthermore, based on our earlier argument, we expect the Treatment × Group interaction effect to be weaker on job-irrelevant scales than on job-relevant scales (Hypothesis 2b).

The third purpose of Study 2 was to investigate how treatment and group membership (flagged vs. nonflagged) would influence applicant perceptions. We expect a Treatment × Group interaction effect, based on the social exchange theory (Blau, 1964), research on self-serving attributions (e.g., Bradley, 1978), and the notion of justice as a means of preserving self-esteem (e.g., Schroth & Shah, 2000). Specifically, when applicants showed strong faking tendencies and got away with it (not warned), they would consider the testing procedure fair, and might even report inflated perceptions as a way to “return the favor.” In this case, their self-esteem remained intact. However, when applicants showing strong faking tendencies were warned, their self-esteem was under attack, and it would be in their best interest from a self-esteem preservation standpoint to blame the testing procedure as being unfair. For nonflagged and unwarned applicants, we do not expect them to inflate their ratings of applicant perceptions, because they did not benefit and hence had no need to “return the favor.” Furthermore, we expect the warning to have a weaker effect on self-esteem for the nonflagged, because the majority of nonflagged applicants were presumable responding honestly, and thus were less concerned about protecting their self-esteem. Note that although the carefully designed warning message should mitigate the negative effect of the warning in Study 2 as it did in Study 1, such an effect should occur among both flagged and nonflagged applicants in Study 2, thus likely canceling each other out. Therefore, we hypothesize that applicants receiving the warning message should report less positive perceptions than applicants receiving the control message; however, the warning effect should be weaker...
among nonflagged applicants than among flagged applicants (Hypothesis 3).

Finally, a secondary goal of Study 2 was to explore the external validity of the proposed procedure. Whereas Study 1 was conducted in a controlled environment, Study 2 was conducted in an unproctored environment. Given the increased usage of unproctored online testing, it is useful to find out the extent to which results of Cell b (flagged and warned) and Cell c (nonflagged and unwarned) may generalize across testing contexts (from Study 1 to Study 2).

Method

Sample and procedure. Participants were college students applying for the membership of an extremely popular student organization, “Student Career Development Association” (SCDA), at the same university where Study 1 was conducted. Due to an increasingly competitive job market in China, job hunting has become a major concern among Chinese college students. The semi-official SCDA was formed several years ago to address students’ needs by offering a variety of career-related services. SCDA members help organize these services and can gain several notable additional benefits such as the opportunities to hone their teamwork and leadership skills, to expand their social network, and to receive numerous internal career-related trainings. Each year, SCDA selects around 100 new members out of roughly 400–500 applications. Résumés and interviews have been used to make selection decisions.

The present research team worked with SCDA to have the Chinese 16PF incorporated into the application process. Applicants were asked to complete a 30-min online psychological assessment as part of the application process. However, unknown to the applicants, the personality test was used for research purposes only, not for actual selection decisions. The testing procedure was the same as Study 1, except that in the present study, the warning and control messages were randomly assigned to all applicants after the initial block. The recruitment period lasted for 2 weeks. After the recruitment had ended, all applicants were carefully debriefed—they were explained the purpose of the study and the necessity of randomly assigning messages; they were assured that interview invitations were made based solely on their résumés; each participant was given a cash reward of 50 Chinese Yuan (approximately $7.84 U.S.) for their time and a short report of their personality profiles.

Three hundred ninety-seven applicants completed at least part of the online test, with 386 providing complete personality data. A chi-square test with Yate’s correction indicated that the attrition rate did not differ across experimental conditions, \( \chi^2(1, N = 11) = 0.14, p = .71 \). Among the 386 applicants, 47.7% were female, 78.2% were undergraduate students, the average age was 20.4 years old, and the fields of study were very diverse. The cell sizes were as follows: nonflagged and unwarned (\( n = 163 \)), nonflagged and warned (\( n = 170 \)), flagged and unwarned (\( n = 30 \)), and flagged and warned (\( n = 23 \)). Thirty-seven out of 386 applicants did not complete any of the optional perception measures. A chi-square test with Yate’s correction revealed no significant difference in the nonresponse rates across experimental conditions, \( \chi^2(1, N = 37) = 0.27, p = .60 \).

Measures. Identical measures as were used in Study 1 were used in Study 2. However, a separate relevance and direction rating for the SCDA position was conducted. Four psychology graduate students at this university familiar with the Chinese 16PF and two senior SCDA members in charge of new member recruitment rated the relevance and favorable direction of Chinese 16PF dimensions to the SCDA position. These SMEs had no knowledge of the proposed procedure. Exactly the same rating procedure used in Study 1 was followed in Study 2. Table 3 lists the averaged relevance ratings and the aggregated direction ratings for 16PF dimensions.

Results

Treatment \( \times \) Group \( \times \) Time interactions. Table 3 presents 20% winsorized descriptive statistics based on the winsorized sample and standardized mean differences over time (within-subject \( d_s \)) in four experimental conditions. Like in Study 1, we used the robust ANOVAs to test the hypotheses. We also conducted Yuen’s robust \( t \) tests in each experimental condition separately.

For the IM scale, the three-way interaction was significant, \( F(1, 382) = 6.21, p < .05, \) partial \( \eta^2 = .02 \). As can be seen in Table 3, on the one hand, the two unwarned groups had very little score change over time. On the other hand, the two warned groups significantly lowered their IM scores over time; however, the flagged and warned group had a larger amount of score reduction than the nonflagged and warned group (\( d = 2.51 \) and 1.59, respectively). Thus, Hypothesis 1a was supported for the IM scale.

For the BS scale, the three-way interaction was significant, \( F(1, 382) = 4.39, p < .05, \) partial \( \eta^2 = .01 \). The two unwarned groups did not change their BS scores much over time. However, the two warned groups lowered their BS scores over time, with the flagged and warned group having a larger amount of score reduction than the nonflagged and unwarned group (\( d = 0.65 \) and 0.12, respectively). The interaction form was similar to that of the IM scale. Thus, Hypothesis 1a was supported for the BS scale.

For the Factor C Emotional Stability scale, the three-way interaction was not significant, \( F(1, 382) = 0.49, p = .49, \) partial \( \eta^2 = .001 \). For the Factor G Rule-Consciousness scale, the three-way interaction was not significant, either, \( F(1, 382) = 2.98, p = .09, \) partial \( \eta^2 = .01 \). Thus, Hypothesis 1a was not supported for the Factors C and G scales. For the Factor N Privateness scale, the three-way interaction was not significant, \( F(1, 382) = 0.10, p = .75, \) partial \( \eta^2 = .0003 \). Comparisons of the three-way interaction effect sizes among the five initial block scales clearly indicated that the interaction was the weakest for Factor N Privateness. Thus, Hypothesis 1b received some support.

For comparison purposes, we repeated the above analyses using the regular three-way mixed design ANOVA. Results based on the regular ANOVA led to the same statistical conclusions except for the BS scale. For this scale, the three-way interaction was not significant, \( F(1, 382) = 2.52, p = .11, \) partial \( \eta^2 = .01 \). Thus, were the regular ANOVA used, Hypothesis 1a would have received somewhat weaker support.

Treatment \( \times \) Group interactions. For the sake of brevity, we report the robust two-way ANOVA results involving non-rettested Chinese 16PF scales and applicant perceptions in a concise way. The full details of robust ANOVA results are available
Table 3
Winsorized Means, Winsorized (Standard Deviations), Standardized Differences, and SME Ratings in Study 2

<table>
<thead>
<tr>
<th>Scale</th>
<th>Flagged and unwarned (n = 30)</th>
<th>Flagged and warned (n = 23)</th>
<th>Nonflagged and unwarned (n = 163)</th>
<th>Nonflagged and warned (n = 170)</th>
<th>Ratings</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Time 1</td>
<td>Time 2</td>
<td>d</td>
<td>Time 1</td>
<td>Time 2</td>
</tr>
<tr>
<td>Impression Management</td>
<td>17.56 (0.91)</td>
<td>17.72 (0.91)</td>
<td>−0.23</td>
<td>17.53 (1.06)</td>
<td>10.07 (5.39)</td>
</tr>
<tr>
<td>Bogus Statement</td>
<td>0.56 (0.87)</td>
<td>0.61 (0.87)</td>
<td>−0.28</td>
<td>0.53 (0.93)</td>
<td>0.13 (0.84)</td>
</tr>
<tr>
<td>Factor A Warmth</td>
<td>14.72 (1.61)</td>
<td>12.40 (2.39)</td>
<td>2.32</td>
<td>11.20 (2.63)</td>
<td>11.82 (2.28)</td>
</tr>
<tr>
<td>Factor B Emotional Stability</td>
<td>19.83 (1.67)</td>
<td>19.78 (1.31)</td>
<td>0.05</td>
<td>20.13 (0.95)</td>
<td>17.07 (1.94)</td>
</tr>
<tr>
<td>Factor C Emotional Stability</td>
<td>15.28 (1.97)</td>
<td>11.93 (2.51)</td>
<td>3.34</td>
<td>13.82 (2.31)</td>
<td>13.67 (2.31)</td>
</tr>
<tr>
<td>Factor D Warmth</td>
<td>19.72 (2.49)</td>
<td>15.93 (2.75)</td>
<td>3.79</td>
<td>16.47 (3.37)</td>
<td>16.45 (3.31)</td>
</tr>
<tr>
<td>Factor E Dominance</td>
<td>16.28 (1.33)</td>
<td>16.28 (1.58)</td>
<td>0</td>
<td>16.60 (1.34)</td>
<td>13.20 (2.03)</td>
</tr>
<tr>
<td>Factor F Liveliness</td>
<td>14.11 (1.62)</td>
<td>14.07 (0.88)</td>
<td>0.04</td>
<td>13.79 (2.26)</td>
<td>14.40 (2.03)</td>
</tr>
<tr>
<td>Factor G Rule-Consciousness</td>
<td>8.83 (1.63)</td>
<td>8.56 (1.33)</td>
<td>0.12</td>
<td>8.87 (1.59)</td>
<td>8.80 (1.79)</td>
</tr>
<tr>
<td>Factor H Social Boldness</td>
<td>12.67 (1.35)</td>
<td>12.53 (1.24)</td>
<td>0.14</td>
<td>11.66 (1.25)</td>
<td>11.27 (1.53)</td>
</tr>
<tr>
<td>Factor I Sensitivity</td>
<td>9.33 (1.63)</td>
<td>8.67 (1.23)</td>
<td>0.66</td>
<td>9.78 (1.60)</td>
<td>8.91 (2.20)</td>
</tr>
<tr>
<td>Factor J Vigilance</td>
<td>15.28 (1.27)</td>
<td>11.33 (2.26)</td>
<td>3.94</td>
<td>12.76 (2.13)</td>
<td>11.23 (1.62)</td>
</tr>
<tr>
<td>Factor K Abstractionedness</td>
<td>5.78 (2.29)</td>
<td>11.00 (1.92)</td>
<td>−5.22</td>
<td>9.49 (2.38)</td>
<td>11.67 (2.62)</td>
</tr>
</tbody>
</table>

Note. The degree of winsorizing is 20%. All d's were within-subject d's. For effect size d's, positive values indicate mean scores were higher in Time 1 than mean scores in Time 2. SME = subject matter expert; R = relevance; D = direction; "+" means higher amount of the trait is favorable; "−" lower amount of the trait is favorable; "?" means direction is unclear.

* p < .05. ** p < .01.
from the authors. We also note that both the robust and regular ANOVAs led to the same statistical conclusions for these scales. Factor A Warmth, Factor F Liveliness, Factor H Social Boldness, Factor L Vigilance, Factor O Apprehension, Factor Q3 Perfectionism, and Factor Q4 Tension all had a high relevance rating and a clear favorable direction. Hypothesis 2a predicted that for these scales, applicants receiving the warning message should have less favorable scores than applicants receiving the control message, but the warning effect should be smaller among nonflagged applicants than among flagged applicants. Robust two-way ANOVAs reveal that the Treatment × Group interaction was significant for all the above scales, except for the Factor O Apprehension scale. Take Factor H Social Boldness as an example: The interaction $F(1, 29) = 9.74, p < .01$, partial $\eta^2 = .25$. Simple effect analysis and Figure 3 show that the warning effect was significant among the flagged and the nonflagged; however, it was larger among the flagged than among the nonflagged ($d = 2.00$ and 0.53, respectively). As for the Factor O Apprehension scale, despite a nonsignificant interaction effect, $F(1, 32) = 2.77, p = .11$, partial $\eta^2 = .08$, the simple effect pattern was consistent with our prediction. Thus, taken together, Hypothesis 2a was largely supported.

Factor E Dominance, Factor I Sensitivity, Factor M Abstractedness, Factor Q1 Openness to Change, and Factor Q2 Self-Reliance all had a low relevance rating and/or an unclear favorable direction. Hypothesis 2b predicted weaker Treatment × Group interaction effects for these scales. Results of robust two-way ANOVAs indicate that the interaction was not significant for four of the above five scales (except for the Factor I sensitivity scale), with effect sizes, partial $\eta^2$, ranging from .0002 to .07 (see Table 4). These effect sizes were smaller than those of the aforementioned highly job-relevant nonretested scales. As for the Factor I Sensitivity scale, the interaction effect was significant, $F(1, 34) = 5.07, p < .05$, partial $\eta^2 = .13$. Simple effect analysis reveals that among the flagged, the warning was associated with lower sensitivity mean scores (11.83 vs. 10.73, $p = .20$, $d = 0.58$), whereas among the nonflagged, the warning was associated with a higher mean sensitivity score (10.77 vs. 11.72, $p < .01$, $d = -0.53$). This interaction pattern was unexpected. Nevertheless, taken together, Hypothesis 2b received partial support.

For the perceived test fairness scale, the Treatment × Group interaction was significant, $F(1, 38) = 6.74, p < .05$, partial $\eta^2 = .15$. Simple effect analysis indicates that unwarned applicants had more positive test fairness perceptions than warned applicants; however, the warning effect was larger among the flagged than among the nonflagged ($d = 1.28$ and 0.19, respectively). For the Perceived Face Validity scale, the interaction was not significant, $F(1, 30) = 0.33, p = .57$ partial $\eta^2 = .01$. For the Test Satisfaction scale, the interaction was not significant, $F(1, 22) = 3.37, p = .08$, partial $\eta^2 = .13$. Although simple effect analyses indicate that for these two scales the interaction tendencies were consistent with Hypothesis 3, they failed to reach statistical significance. For the Test-Taking Motivation scale, the interaction was not significant, $F(1, 25) = 0.03, p = .86$, partial $\eta^2 = .00004$, showing no predicted interaction trend. As can be seen in Table 3, flagged and unwarned applicants tended to report more positive perceptions than nonflagged and unwarned applicants. A series of Yuen’s $t$ tests indicate that the group differences in perceived test fairness and perceived face validity were significant. Thus, Hypothesis 3 was supported for the perceived test fairness scale, but was not supported for the other three perception scales.

**External validity.** We compared results of Cells b and c across the two studies. It turned out that although effect sizes varied somewhat between the two studies, the result patterns (e.g., Treatment × Group interactions for initial block scales, and group mean differences in nonretested job-relevant scales and applicant perceptions scales) were quite similar. Thus, the proposed procedure demonstrated reasonable external validity across testing contexts. The full details of these comparisons are available from the authors.

**Discussion**

The results of Study 2 provided additional insights into the proposed procedure. Internal validity was established by showing that the warning was responsible for the score reduction of flagged applicants. However, the evidence for construct validity was mixed. The hypothesized three-way interaction was supported for the IM and BS scales, but not for the two highly job-relevant 16PF scales (Factors C and G). At first glance, these null findings pose a serious threat to construct validity. Fortunately, for nonretested job-relevant 16PF scales, we found the hypothesized interaction effect quite consistently.

How can we reconcile these seemingly contradictory findings? We speculate that the nonsignificant three-way interactions on the Factors C and G scales were likely due to applicants somehow recognizing those retested items showing up again on the screen, which prompted them to respond to the experimental demand on those items. In contrast, for nonretested scales, items were presented only once, thus not subject to the above influence. It seems that the warning effect, not the experimental demand effect, carried over to the nonretested scales in the main block. Significant three-way interactions on the IM and BS scales were most likely due to the fact that scores on these two scales were used to
determine group membership (flagged vs. nonflagged). Consequently, flagged applicants had much more “space” to reduce their scores on these two scales than nonflagged applicants—by contrast, 16PF scales were less prone to this effect. Future research is needed to explore ways to alleviate the experimental demand effect on responses to retested job-relevant 16PF scales, an issue to which we return subsequently.

As predicted, the Treatment × Group interaction effect was weaker for many job-irrelevant 16PF scales. These findings continue to support the sophisticated view of job applicants (e.g., Rosse et al., 1998; Vasilopoulos et al., 2005). That is, job applicants are quite deliberate in deciding whether and how to fake on personality items. The unexpected significant Treatment × Group interaction effect on the Factor I Sensitivity scale might reflect job applicants’ unfamiliarity with some of the nuances of the target position and the resulting incorrect judgments they formed regarding the job relevance of certain personality items.

With respect to applicant perceptions, there was interesting, yet somewhat mixed evidence suggesting that potential fakers (a) tended to inflate their ratings of fairness-related perceptions, relative to nonfakers, and (b) reacted more negatively to the warning. Thus, faking appears to be a potentially important variable that has seldom been looked at, but which warrants further research in the applicant perceptions literature. The null finding for test-taking motivation was not surprising, because attitude and cognition perceptions are generally considered more relevant in cognitive test contexts than in noncognitive test contexts (Ryan & Ployhart, 2000).

General Discussion

In the present research, we proposed a new procedure for reducing faking on personality tests. Results based on two field studies showed some promises for the proposed procedure. One important theoretical implication is that the present research helps clarify the faking construct. We conceptualized faking as a deliberate attempt to tailor one’s test responses to the demands of a particular testing situation. The results of our study demonstrate that applicants do indeed engage in faking and that levels of faking were reduced (although not completely eliminated) after applicants whose scores on an IM or a BS scale suggested that they were receiving a targeted warning. Taken together, these findings provide support for our conceptualization of faking as intentional, situationally induced, and changeable. Moreover, these findings lend credence to the notion that faking is indeed a cause for concern (e.g., Landers et al., 2011), contrary to what some researchers have suggested (e.g., Hogan, Barrett, & Hogan, 2007).

As we noted earlier, the use of warnings is not a new idea, as research has shown that traditional warnings can reduce faking on personality tests (Dwight & Donovan, 2003). Compared with traditional warnings, which entail simply giving all applicants a verbal or written warning before the test, the proposed procedure may seem more complicated, as it requires (a) developing valid measures of faking, (b) determining appropriate cut scores, (c) designing a polite warning message, and (d) delivering the warning to those who exceed the cut scores. Thus, it is reasonable to question whether this extra work is worth it.

We believe that the proposed procedure has a number of benefits over traditional warnings. Philosophically, traditional warnings implicitly assume that most, if not all, applicants would fake, hence the need to give all applicants an initial warning. However, such an assumption runs counter to recent empirical findings that a substantial percentage of applicants do not fake (e.g., Donovan, Dwight, & Hurtz, 2003; Griffith et al., 2007). By contrast, the proposed procedure approaches applicants from a tabula rasa

<table>
<thead>
<tr>
<th>Scale</th>
<th>Robust interaction effect</th>
<th>Unwarned vs. warned effect size (d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>16PF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Factor A Warmth 8.28 (1, 31) &lt;.01 0.21</td>
<td>1.16* 0.25</td>
<td></td>
</tr>
<tr>
<td>Factor E Dominance 3.06 (1, 40) .09 0.07</td>
<td>1.50** 0.56*</td>
<td></td>
</tr>
<tr>
<td>Factor F Liveliness 8.39 (1, 41) &lt;.01 0.17</td>
<td>1.44** 0.01</td>
<td></td>
</tr>
<tr>
<td>Factor H Social Boldness 9.74 (1, 29) &lt;.01 0.25</td>
<td>2.00** 0.53**</td>
<td></td>
</tr>
<tr>
<td>Factor I Sensitivity 5.07 (1, 34) &lt;.05 0.13</td>
<td>0.58 0.53**</td>
<td></td>
</tr>
<tr>
<td>Factor L Vigilance 4.58 (1, 34) &lt;.05 0.12</td>
<td>-1.87** 0.69**</td>
<td></td>
</tr>
<tr>
<td>Factor M Abstractedness 0.90 (1, 51) .35 0.02</td>
<td>0.03 0.28</td>
<td></td>
</tr>
<tr>
<td>Factor O Apprehension 2.77 (1, 32) .11 0.08</td>
<td>-2.18** 0.18**</td>
<td></td>
</tr>
<tr>
<td>Factor Q1 Openness to Change 0.16 (1, 38) .70 0.004</td>
<td>0.11 0.28</td>
<td></td>
</tr>
<tr>
<td>Factor Q2 Self-Reliance 0.07 (1, 29) .79 0.002</td>
<td>0.46 0.46</td>
<td></td>
</tr>
<tr>
<td>Factor Q3 Perfectionism 7.71 (1, 27) &lt;.05 0.22</td>
<td>2.24** 1.07**</td>
<td></td>
</tr>
<tr>
<td>Factor Q4 Tension 8.54 (1, 35) &lt;.01 0.20</td>
<td>-2.48** 0.07**</td>
<td></td>
</tr>
</tbody>
</table>

Applicant perceptions

<table>
<thead>
<tr>
<th>Scale</th>
<th>Robust interaction effect</th>
<th>Unwarned vs. warned effect size (d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perceived Test Fairness 6.74 (1, 38) &lt;.05 0.15</td>
<td>1.28** 0.19</td>
<td></td>
</tr>
<tr>
<td>Perceived Face Validity 0.33 (1, 30) .57 0.01</td>
<td>0.86 0.47</td>
<td></td>
</tr>
<tr>
<td>Test Satisfaction 3.37 (1, 22) .08 0.13</td>
<td>1.13* 0.19</td>
<td></td>
</tr>
<tr>
<td>Test-Taking Motivation 0.001 (1, 25) &lt;.05 0.02</td>
<td>0.57 0.31</td>
<td></td>
</tr>
</tbody>
</table>

Note. Positive effect sizes (d) indicate that mean scores of the unwarned applicants were higher than mean scores of the warned applicants.

*p < .05. **p < .01.
standpoint, allowing them to fully express themselves on the test, and does not warn them unless there is compelling evidence of faking. Such a stance is not only more ethically defensible, it is also more consistent with the notion of evidence-based practice (e.g., Rousseau, 2006).

If traditional warnings stop at simply giving all applicants a warning before the test, the hiring organization is left with the challenging issue of how to deal with potential fakers after the test has been completed. Ignoring potential fakers would result in penalizing honest applicants, and ironically making the warning a bluff that would be easy for applicants to circumvent, especially with coaching. However, the effective management of potential fakers requires at least (a) developing valid measures of faking and (b) determining appropriate cut scores, both of which are features of the proposed procedure. Thus, when the entire testing and selection processes are considered, the proposed procedure is not necessarily substantially more complicated than traditional warnings.

One real advantage of the proposed procedure is that it attempts to manage potential fakers proactively during (rather than after) the testing process. It holds potential fakers accountable through sending them a clear message that potential fakers are identified and warned. When dealing with defiant fakers after the test, the proposed procedure also affords the hiring organization greater certainty for punitive actions, relative to traditional warnings. This is because in the proposed procedure, applicants need to exceed the faking criterion twice (in the initial block and the main block) to be classified as defiant fakers, whereas in traditional warnings, defiant fakers are those who exceed the faking criterion once. In the case of Study 1, the hiring organization decided to adopt an even more stringent standard to single out those applicants who exceeded both faking criteria (BS and IM) at both time points. Such a practice is ethically more defensible than relying on just social desirability scores alone and once. That being said, we acknowledge that the proposed procedure does not eliminate the problem of what to do with potential fakers; nevertheless, it reduces the size of the problem and gives the hiring organization a better position to address this issue.

The proposed procedure also has some weaknesses, relative to traditional warnings. For instance, cut scores are chosen on the basis of continuous IM and BS scores. As a result, the proposed procedure might favor milder fakers whose IM and BS scores fall just short of the cut scores. Interestingly, the proposed procedure has a potential construct validity issue, which also plagues traditional warnings, but in a slightly different manner. Specifically, Kuncel and Borneman (2007) correctly noted that one condition for the effectiveness of traditional warnings is that warnings do not suppress the responses of honest applicants. On the other side of the coin, we found in Study 2 that the warning frequently lowered nonflagged applicants’ scores. Although in field settings nonflagged applicants should not be warned, the above findings do not rule out the possibility that the warning could lead to flagged applicants overcorrecting their personality scores. Despite these potential weaknesses, we believe that the proposed procedure at least represents a promising alternative to traditional warnings that warrants further study.

Although the proposed procedure offers some promise, a variety of practical issues need to be explored and considered before being used operationally, particularly in North American countries, where the testing context is often dictated by the legal environment. As an anonymous reviewer pointed out, the proposed procedure might place an organization at risk legally. For instance, the proposed procedure entails sending different types of messages (warning vs. control) to different types of applicants (potential fakers vs. nonfakers) after the initial block. This differential treatment of applicants during the testing process might raise the potential for disparate impact. Furthermore, there is some empirical evidence that certain racial groups (e.g., Asians, Blacks, and Hispanics) tend to score higher on social desirability scales than the White group in actual selection contexts (e.g., Dudley, McFarland, Goodman, Hunt, & Sydell, 2005; Hough, 1998). Certain minority applicants thus might be more likely to be flagged as potential fakers and to receive the warning message. This in turn might result in these minority applicants having significantly lower personality scores as compared with majority applicants (Dudley et al., 2005). In other words, the proposed procedure might unintentionally introduce racial adverse impact. Therefore, we call for future research to systematically examine whether the proposed procedure may influence different groups differentially. Until evidence is accumulated to demonstrate that the procedure does not result in adverse impact, we advise using caution when implementing the new procedure in North American organizations.

Another issue concerns whether our results obtained in China will generalize to the U.S. and other Western cultures. For instance, Chinese culture is generally considered a collectivistic and high power-distance culture (Markus & Kitayama, 1991). Chinese people tend to value group harmony and respect for authority. Thus, Chinese job applicants, after receiving the warning message from an authoritative source (the hiring organization), should willingly lower their personality scores and at the same time should not develop very negative perceptions about their testing experience, and are not likely to quit the personality test. By contrast, the United States, like many other Western cultures, is generally considered an individualistic and low power-distance culture. American people tend to value independence, fairness, and autonomy. Thus, American job applicants, after receiving the warning message, may be less willing to lower their personality scores, may develop negative perceptions about their testing experiences, and may be more likely to quit the personality test. If this is true, one may expect that the warning effect on personality scores should be larger in Chinese culture than in U.S. culture, whereas the warning effect on applicant perceptions should be smaller in Chinese culture than in U.S. culture. We encourage future researchers to test the proposed procedure across diverse cultures.

**Study Limitations and Future Research Directions**

The present research has some limitations. The first limitation was that the warning message was longer than the control message, thus creating a potential confound. Our decision to use the current sets of messages in the present research was primarily driven by practical considerations. Future researchers should develop and test control messages that are similar in length to the warning message to address the confounding issue.

The second limitation was that the contexts of Study 1 and Study 2 were quite different in terms of type of sample (job applicants vs. student organization applicants), the testing envi-
vironment (proctored vs. unproctored), and the consequences of the testing itself (whether getting the real job vs. whether getting into the student organization). Supporting this point, Study 2 had a lower percentage of applicants being flagged as potential fakers than Study 1 (14% vs. 29%); and Study 2 applicants tended to have less favorable personality scores than Study 1 applicants (see Tables 1 and 3). These differences might have rendered the results of the two studies not directly comparable. Therefore, we call for future research to replicate Study 2 findings in actual employee selection contexts.

Future research is needed to further examine the construct validity of the proposed procedure. The key issue is to ensure that the warning does not lead applicants (particularly flagged applicants) to overcorrect their scores. To fully address construct validity, a more sophisticated design shall be used, which requires (a) having a high selection ratio to avoid subject attrition and (b) having accepted applicants complete the main block items again in a low-motivation context, for example, after organizational entry. Then the researcher can compare applicants’ personality scores in the initial and main blocks, and postentry, to gauge the extent to which applicants overcorrect their scores after being warned. If construct validity continues to present a problem in these studies, scholars may consider several strategies to address it, for instance, softening the tone of the warning message by removing the consequent part and/or including all job-irrelevant scales (except for the IM and BS scales) in the initial block.

Future research is also needed to examine whether the proposed procedure can improve the criterion-related validity of personality scores, particularly among flagged applicants. When job performance data are available, we may compare the criterion-related validity using initial block personality scores versus main block personality scores as predictors. Showing improved criterion-related validity among flagged applicants will be a critical piece of evidence for gaining selection professionals’ acceptance of the proposed procedure.

Vasilopoulos et al. (2005) recently showed that traditional warnings may increase the correlations between personality scores and cognitive ability scores, thus indirectly causing adverse impact. It will be interesting to investigate whether such a tendency may also occur in the proposed procedure, in which the warning was sent out during (rather than before) the test. It is possible that different timings of the warning might evoke different psychological mechanisms, resulting in the proposed procedure being more or less prone to the above undesirable tendency.

Future research should also compare the relative effectiveness of the proposed procedure with other existing faking-mitigation methods such as traditional warnings. Finally, future research should test the proposed procedure using other personality inventories such as the California Psychological Inventory, Hogan Personality Inventory, and NEO Personality Inventory. With more programmatic research to further refine and test the proposed procedure, we hope that researchers and practitioners will eventually be provided with a useful tool to effectively manage applicant faking on personality tests within selection contexts.

References


Received June 18, 2010
Revision received August 26, 2011
Accepted November 14, 2011