Using Blatant Extreme Responding for Detecting Faking in High-stakes Selection: Construct validity, relationship with general mental ability, and subgroup differences

Julia Levashina*, Jeff A. Weekley, Nicolas Roulin*** and Erica Hauck****

*College of Business Administration, Kent State University, Kent, OH 44242-0001, USA. jlevashi@kent.edu
**IBM, Frisco, TX, USA
***University of Manitoba, Winnipeg, MB, Canada

Although there has been a steady growth in research and use of self-report measures of personality in the last 20 years, faking in personality testing remains as a major concern. Blatant extreme responding (BER), which includes endorsing desirable extreme responses (i.e., 1 and 5 s), has recently been identified as a potential faking detection technique. In a large-scale (*N* = 358,033), high-stakes selection context, we investigate the construct validity of BER, the extent to which BER relates to general mental ability, and the extent to which BER differs across jobs, gender, and ethnic groups. We find that BER reflects applicant faking by showing that BER relates to a more established measure of faking, an unlikely virtue (UV) scale, and that applicants score higher than incumbents on BER. BER is (slightly) positively related to general mental ability whereas UV is negatively related to it. Applicants for managerial positions score slightly higher on BER than applicants for non-managerial positions. In addition, there was no gender or racial differences on BER. The implications of these findings for detecting faking in personnel selection are delineated.

International Journal of SELECTION AND ASSESSMENT

1. Introduction

The use of personality inventories in the selection has increased greatly over the last two decades as a means for identifying applicants best suited for a job. However, concerns regarding applicants' ability to fake or intentionally distort responses have accompanied the increased utilization of these tools (Morgeson et al., 2007). Self-report measures are susceptible to being faked by motivated job applicants because the 'right' answer may be apparent to them (Douglas, McDaniel, & Snell, 1996; McFarland & Ryan, 2000; Ryan & Sackett, 1987; Zickar & Drasgow, 1996). When comparing participants instructed to respond honestly to those instructed to fake, a meta-analysis showed that fakers scored 0.5 standard deviations higher than participants who responded honestly (Viswesvaran & Ones, 1999). When comparing applicants to nonapplicants across several jobs, another meta-analysis showed that applicants demonstrate about one third of a standard deviation advantage over nonapplicants on personality measures (Birkeland, Manson, Kisamore, Brannick, & Smith, 2006; Hough, Eaton, Dunnette, Kamp, & McCloy, 1990).

The primary concern with faking is its potential impact on criterion-related validity (Hough et al., 1990). Although some research has shown little to no effect (e.g., Ones & Viswesvaran, 1998; Schmitt & Oswald, 2006), others have demonstrated that faking decreases test validity (e.g., Dunnette, McCartney, Carlson, & Kirchner, 1962; Harold, McFarland, & Weekley, 2006; Ones, Viswesvaran, & Schmidt, 1993). Furthermore, meta-analytic results demonstrate that predictive validities are on average 0.07 points lower than concurrent validities (Huffcutt, Conway, Roth, & Klehe, 2004), suggesting that when applicants are motivated to fake there may be consequences for criterion-related validity. Ad-

9600 Garsington Road, Oxford, OX4 2DQ, UK and 350 Main St., Malden, MA, 02148, USA

ditionally, applicant faking may change the rank order of applicants in a selection setting and thus alter hiring decisions (Donovan, Dwight, & Hurtz, 2003; McFarland & Ryan, 2000; Rosse, Stecher, Miller, & Levin, 1998). Furthermore, fakers are more likely to engage in workplace deviant behavior once hired (O'Neill et al., 2013). Applicant's perception of test fakability may also reduce their reactions and test taking motivation.

As such, researchers and practitioners have explored a variety of methods for detecting and controlling faking, including the use of optimal indexes (Drasgow, Levine, & McLaughlin, 1987), response latencies (Holden & Hibbs, 1995), appropriateness measures (Zickar & Drasgow, 1996), unlikely virtue (UV) scales (Hough, 1998), warnings (Dwight & Donovan, 2003), forced choice scales (Rothstein & Goffin, 2006), idiosyncratic item responses (Kuncel & Borneman, 2007), *Z* tests or likelihood ratios (Guo & Drasgow, 2010), covariance indexes (O'Connell, Kung, & Tristan, 2011), implausible answers (O'Connell et al., 2011), or even eye tracking technology (van Hooft & Born, 2012). The most widely used approach includes using faking or lie scales, which can be categorized into two groups: inventory specific versus general. Inventory-specific scales are designed to detect deception in a particular personality inventory, such as the Hogan Personality Inventory (Hogan, 1986) or the Minnesota Multiphasic Personality Inventory (MMPI; Dahlstrom, Welsh, & Dahlstrom, 1972; Dahlstrom, Welsh, & Dahlstrom, 1975). General scales are developed to be used with any personality inventory and are more widely utilized. Social desirability (SD) scales are an example of general deception detection scales. A survey of personality researchers showed that 56% typically use personality measures with a faking scale (Goffin & Christiansen, 2003). Moreover, 69% were in favor of correcting personality scores based on SD or other faking scales. Yet, recent research has demonstrated that there are more problems than benefits with the use of SD scales, resulting in a call for more research on alternative measures of faking (Griffith & McDaniel, 2006; Griffith & Peterson, 2008; Kuncel & Tellegen, 2009).

Recently identified as a new type of faking, blatant extreme responding (BER, Landers, Sackett, & Tuzinski, 2011) involves endorsing desirable extreme points, such as 1 and 5 s on a 5-point Likert scale, in order to achieve a maximum score on a measure. BER builds on the assumption that applicants responding honestly will be unlikely to endorse extreme responses across many items on a personality measure because people typically possess varying levels of different traits. Two studies have explored BER as a potential new faking detection method (Landers et al., 2011; Weekley & McKee, 2009) and provided initial empirical support for the usefulness of BER. Yet, several issues remain to be addressed.

First, we do not know what BER actually measures. More specifically, applicants' extreme responses may reflect (a) true individual differences (e.g., applicants responded honestly as they indeed possess a very high level of that particular personality trait); (b) a general tendency to use the extreme points when responding to questions (i.e., an extreme response style or ERS; Möttus et al., 2012); and (c) an intentional increase of scores to look like a better fit with the job or organization (i.e., deliberate faking; Landers et al., 2011). Therefore, it is important to examine the construct validity of BER as a faking detection technique. Past research has acknowledged this, calling for further investigation of the construct validity of BER (Landers et al., 2011).

Second, an emerging consensus among researchers is that 'ability to fake' is one of the main components of the faking process (Levashina & Campion, 2006; McFarland & Ryan, 2000; Snell, Sydell, & Lueke, 1999; Tett, Freund, Christiansen, Fox, & Coaster, 2012). Yet, there are inconsistencies in the results of existing studies. Some studies suggest that applicants with high levels of mental ability fake more because they are more capable of recognizing the opportunity to fake (Furnham, 1986; Lao, 2001) and constructing and using a successful faking strategy (Austin, Hofer, Deary, & Eber, 2000; Pauls & Crost, 2004). Other studies suggest that applicants with high levels of mental ability do not fake because they might believe that they can obtain a high score on a measure without faking (De Fruyt, Aluja, García, Rolland, & Jung, 2006). Therefore, it is important to examine the relationship between general mental ability (GMA) and BER.

Third, personality measures typically have the advantage of demonstrating only small differences between ethnic groups, gender, or job type (Hough & Oswald, 2000; Ployhart & Holtz, 2008). Yet, there is some evidence that members of various subgroups fake to a different extent (Weekley, Ployhart, & Harold, 2004). Therefore, respondents from different subgroups may engage differently in extreme responding, making it important to evaluate potential job, gender, and ethnic group differences on BER scores.

Therefore, there are three primary purposes and contributions of this study. First, we examine the construct validity of BER as a faking detection technique. Second, we examine the relationship between BER and GMA. Third, we examine job, gender, and ethnic group differences on BER.

2. Construct validity of BER

BER has recently emerged as a potential new type of faking detection technique (Landers et al., 2011; Weekley & McKee, 2009). BER builds on the assumption that it is very unlikely to expect that applicants who respond honestly will endorse extreme or maximum responses across many items on a personality measure because applicants typically possess varying levels of different traits. As such, there should be variability across trait scores, and exceptionally few should be high on every trait measured. Recently, van Hooft and Born (2012) provided indirect support for this assumption. Using eye tracking technology, they showed that fakers focus more on the extreme points of the scales when completing personality tests. Landers et al. (2011) showed that internal applicants who likely were informally coached to obtain a high score by using only extremes of the response scales retook the test after initial failure, and demonstrated an increase in BER rate and test scores. Weekley and McKee (2009) tested the effect of BER correction on criterion-related validity with mixed results. Three applicant samples were evaluated with only one showing improvements in validity when applicants demonstrating BER were removed from the analysis.

Yet, before one uses BER as a faking detection technique, it is important to examine its construct validity. More specifically, applicants can display extreme responding because of at least three reasons. First, extreme responses may reflect true individual differences. Indeed, it is possible that some applicants actually possess a very high level of that particular personality trait and that extreme responses are honest descriptions of these qualities. Second, extreme responses may reflect a general tendency to use the extreme points when responding. Research on ERS has highlighted that some people tend to systematically use more the extreme points of the scales when responding to surveys, questionnaires, or tests (e.g., Arce-Ferrer, 2006; Cheung & Rensvold, 2000; De Beuckelaer, Weijters, & Rutten, 2010; de Jong, Steenkamp, Fox, & Baumgartner, 2008; Möttus et al., 2012; Wetzel, Carstensen, & Böhnke, 2013). This tendency has been observed in responding to personality inventories, such as the NEO-FFI (Austin, Deary, & Egan, 2006) or the NEO-PI-R (Wetzel et al., 2013). Finally, extreme responses may reflect deliberate faking to look like a better fit with the job or the organization (Landers et al., 2011). Interestingly, although ERS has been extensively discussed in the personality literature, it has been seldom examined in applied settings (McGrath, Mitchell, Kim, & Hough, 2010), and no research has attempted to distinguish ERS from faking (Wetzel et al., 2013).

As such, we propose to examine the construct validity of BER, that is, demonstrate that BER actually captures intentional response distortion and not simply honest high scores on a personality trait or ERS in two ways: (a) by showing that applicants score higher than incumbents on personality traits, BER, and a more established measure of faking, but not on a measure not susceptible to faking (i.e., GMA), and (b) by relating BER to a more established measure of faking.

A first approach to measure faking involves comparing applicant and incumbent responses on a given personality measure. This indirect approach assumes that incumbents should have little motivation to fake, whereas applicants are motivated to get the job and fake to increase their scores on personality measures. Therefore, score differences observed between the two groups reflect intentional distortion and provide indirect evidence of faking (Ellingson, Sackett, & Connelly, 2007; Griffith & Converse, 2011; Ziegler, MacCann, & Roberts, 2011). Past research comparing applicant scores with incumbent scores found applicants to score significantly higher than incumbents (Birkeland et al., 2006). Thus, we expect to find higher scores on personality measures for applicants than incumbents, and this difference should also be reflected in BER scores.

If we observe higher BER scores for applicants than incumbents, we can demonstrate that BER captures faking and we can eliminate truly high personality traits and ERS as explanations for BER. Indeed, it is very unlikely to have more people possessing highly desirable traits in an applicant sample than in an incumbent sample. It is rather more likely to have more people possessing those traits in the *incumbent* sample because they have been selected (at least partly) on those traits to join the organization. Moreover, it is very unlikely to have more people using a systematic ERS in an applicant sample than in an incumbent sample. People using ERS tend to score higher on conscientiousness (and extraversion) (Austin et al., 2006; Möttus et al., 2012). And conscientiousness has been identified as the personality trait the most predictive of job performance (Schmidt & Hunter, 1998). As such, incumbents are likely to be truly more conscientious than applicants, again because they have been selected (at least partly) on those traits to join the organization, and thus to use more ERS.

In addition, if differences between applicants and incumbents (i.e., on personality scores and BER) are due to intentional distortion (and not other differences in sample compositions), we should also observe such difference on another measure of faking, such as UV scales (Hough, 1998; Hough et al., 1990; Hurtz & Alliger, 2002). UV scales are designed to measure behaviors that are implausibly virtuous (e.g., 'I have never been grouchy with anyone') so that they are unlikely to describe anyone. Thus, when applicants endorse such items, they are believed to be intentionally distorting their responses in order to make a good impression or to increase their scores on a personality measure. Hough (1998) reported higher UV scores for applicants than incumbents in two of three samples. There is also evidence that UV corrections may identify a subset of individuals who *should* have been hired, as well as those who *would not* have been hired, thus improving hiring decisions (Hough, 1998; Weekley, 2006). Conversely, we should not observe any difference between the applicant and incumbent sample on measures that are not susceptible to faking, such as GMA tests (e.g., Arthur, Glaze, Villado, & Taylor, 2010).

Second, if BER measures faking, then it should demonstrate convergent validity with other measures of faking (Cronbach & Meehl, 1955), such as a direct measure intentional distortion (i.e., UV). Thus, we expect that applicants who endorse UV items should also evidence BER. Yet, this relationship should be weaker in a nonmotivated sample, such as with incumbents, than with a motivated sample, such as with applicants. Based on the above rationale, we propose the following hypotheses to examine the construct validity of BER:

Hypothesis 1: Applicants will score higher than incumbents on (a) personality, (b) BER, and (c) UV, but (d) not on GMA.

Hypothesis 2: BER will be positively related to UV in both applicant and incumbent samples, and the relationship will be stronger in the applicant sample.

3. Relationship between GMA and BER

Research suggests that GMA may influence faking behavior. When applicants are motivated to fake, those high in GMA appear to be better at faking (Egan, 1989; Lao, 2001; Levashina, Morgeson, & Campion, 2009; Mersman & Shultz, 1998; Ones, Viswesvaran, & Reiss, 1996; Pauls & Crost, 2004). Applicants with high levels of GMA can better define job-relevant personal characteristics and therefore are better able to respond to personality measure in a manner consistent with a qualified job applicant (Furnham, 1990). Therefore, we should observe a stronger positive relationship between GMA and personality scores in the applicant context than in the incumbent context.

BER results from consistently endorsing the extremely desirable response options. Although it is not a cognitively complex strategy to improve one's test score, it still requires applicants to identify the desirable end of the scale for each item. Applicants with high levels of GMA are likely to be cleverer test takers who endorse items that contribute to BER. Therefore, BER should be positively (but not strongly) related to GMA in the applicant sample. Alternatively, UV results from endorsing implausibly virtuous items. Past research found a negative correlation between GMA and UV (Pulakos et al., 2002; Weekley, 2006), suggesting that applicants high in GMA will be less likely to endorse (or deliberately fail to endorse) UV items than those low in GMA.

Hypothesis 3: GMA will correlate significantly higher with personality in the applicant than in the incumbent sample.

Hypothesis 4: GMA will positively correlate with BER but will negatively correlate with UV in the applicant sample.

4. Job type and subgroup differences on BER scores

Because BER is a general measure of faking, it may be used for a large variety of jobs, including managerial and nonmanagerial positions. Landers et al. (2011) only tested BER in managerial jobs, and thus research examining job type differences for BER is still lacking. Similarly, Hough (1998) only tested UV in nonmanagerial jobs. In their meta-analysis, Birkeland et al. (2006) found differences in faking between managerial and nonmanagerial jobs only for conscientiousness (with more faking for nonmanagerial jobs), but they did not examine BER or UV measures. Moreover, past research has demonstrated that faking varies by job requirements. Applicants fake personality traits that they perceive to be job relevant (Furnham, 1990; Tett et al., 2012). As such, if the personality measures are equally job relevant for managerial and nonmanagerial positions, we should not observe any difference in faking (i.e., in BER and UV scores) between the two samples. Any potential differences would suggest that BER or UV works differently depending on job type. Based on the limited theoretical background and empirical evidence available, we propose to examine potential job type differences in BER and UV scores with the following research question:

Research Question 1: Do incumbents and applicants for managerial and nonmanagerial jobs receive different scores on BER and UV?

Past research suggests that there are some gender and race differences on SD scales, with men scoring higher on SD scales than women (Ones & Viswesvaran, 1998), and Whites scoring lower than minority groups (Hough, 1998; Hough, Oswald, & Ployhart, 2001). For instance, Dudley, McFarland, Goodman, Hunt, and Sydell (2005) showed that Whites scored lower on an SD scale than Blacks ($d = 0.37$), Hispanic ($d = 0.47$), and Asians ($d = 1.04$). The race differences were found across applicant and incumbent context, suggesting that the differences were not situation dependent. Yet, these differences were not observed on the personality measures, suggesting that fewer minority applicants would be selected if SD scales were used to correct personality test scores for SD responding.

To date, there has been no empirical investigation of subgroup differences in BER scores. Moreover, evidence regarding UV-subgroup differences is somewhat mixed. Whereas Hough (1998) found trivial differences, Weekley (2006) reported moderate to large demographic differences on the UV scale, with non-Whites and females scoring higher on both the UV scale and the personality composite than Whites or males. As such, we propose to examine potential demographic differences using the following research questions:

Research Question 2: Are there significant (a) racial and (b) gender subgroup differences on BER and UV scores?

5. Method

5.1. Participants

Two different samples were included in this study. The first sample included 2,295 incumbents from three national retail chains. Of these, 49.3% had a managerial position (i.e., store and assistant store managers) and 51.7% were hourly workers, 50.8% were women and 49.2% men (107 incumbents were missing gender data), 70.1% were White, 13.7% Black, 8.8% Hispanic, 7% Asian, and 0.5% other ethnic backgrounds (536 incumbents were missing race data), and the average age was 35.3 ($SD = 10.4$, age was available from 1,318 incumbents). Because these data were collected for research purposes only and individual responses were not shared with the sponsoring organizations, participants should have been responding in a relatively honest way.

The second sample included a total of 355,738 applicants for positions with a national retailer, a national logistics service company, or a hotel operator. In each case, the assessment comprised one aspect of the employment process determining whether or not an offer was extended. Of these, 8.2% applied for a managerial position and 91.8% for a nonmanagerial position, 49.2% were women and 50.8% men (132,502 applicants were missing gender data), 45.8% were White, 35.1% Black, 13.9% Hispanic, 3.7% Asian, and 1.5% other ethnic backgrounds (166,844 applicants were missing race data), and the average age was 35.2 ($SD = 10.8$, age was available from 4,932 applicants).

5.2. Measures

5.2.1. Personality

The personality measure used in this study was a part of larger inventory developed to reflect work styles taxonomy embedded in the O*NET content model (Peterson et al., 2001; Weekley, Ployhart, & Cooper-Hakim, 2005). Incumbents completed 8–10 measures of personality traits from the 16 in the original O*NET work styles taxonomy. Applicants completed three to four measures. A job analysis and concurrent validation study were conducted to select measures that were relevant to each job. Each scale included 10 items rated on a 5-point scale (*strongly agree* to *strongly disagree*). A complete list of the personality measures and their reliability coefficients can be found in Table 1.

Table 1. Means, standard deviations, and correlations among study variables

	M	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1. General mental ability	0.58	0.23	(0.87)																	
2. Achievement	4.66	0.47	0.07	(0.85)																
3. Initiative	4.63	0.49	0.09	0.80	(0.87)															
4. Adaptability	4.05	0.54	0.01	0.51	0.51	(0.78)														
5. Attention to detail	4.48	0.53	0.03	0.53	0.70	0.37	(0.81)													
6. Energy	4.52	0.52	0.08	0.67	0.78	0.51	0.42	(0.85)												
7. Concern for others	4.11	0.61	0.10	0.32	0.34	0.29	0.26	0.18	(0.71)											
8. Cooperation	4.75	0.46	−0.05	0.87	0.82	0.49	0.43	0.75	0.63	(0.91)										
9. Dependability	4.29	0.53	0.00	0.57	0.73	0.44	0.65	0.50	0.32	0.53	(0.81)									
10. Self-control	4.42	0.54	0.01	0.37	0.64	0.51	0.58	0.68	0.26	0.66	0.45	(0.78)								
11. Social orientation	4.42	0.46	0.13	0.59	0.59	0.45	0.33	0.41	0.57	0.69	0.40	0.38	(0.76)							
12. Stress tolerance	3.95	0.67	0.02	0.45	0.48	0.62	0.40	0.57	0.20	0.42	0.48	0.73	0.39	(0.85)						
13. Integrity	3.72	0.63	−0.16	0.39	0.41	0.31	0.31	0.34	0.34	0.47	0.44	0.40	0.40	0.36	(0.72)					
14. Persistence	4.49	0.48	0.10	0.67	0.75	0.58	0.62	0.68	0.30	0.53	0.69	0.51	0.43	0.60	0.41	(0.77)				
15. Leadership	4.47	0.51	0.11	0.64	0.67	0.52	0.39	0.59	0.19	0.41	0.42	0.36	0.55	0.50	0.36	0.60	(0.84)			
16. Persuasion	3.83	0.61	0.15	0.40	0.42	–	–	0.40	–	–	–	–	–	–	–	0.37	0.53	(0.80)		
17. Analytical thinking	4.44	0.54	0.17	–	0.74	–	–	–	–	–	–	–	–	–	–	0.71	–	–	(0.85)	
18. Unlikely virtue	3.13	1.07	−0.13	0.20	0.27	0.05	0.35	0.33	−0.08	0.23	0.26	0.34	0.12	0.16	0.32	0.33	0.15	0.20	0.31	(0.92)
19. Overall BER	0.70	0.27	0.06	0.68	0.76	0.63	0.73	0.80	0.48	0.71	0.67	0.74	0.68	0.63	0.54	0.78	0.78	0.58	0.80	0.39

Note: All values above .05 are significant at $p < .001$. Reliability coefficients are presented in diagonal (in parentheses). BER = blatant extreme responding. *BER overall* represents BER across all personality subscales. The reliability coefficients are similar across both incumbent and applicant samples.

5.2.2. General mental ability

A 32-item measure of GMA was used. Consisting of 16 numerical and 16 verbal reasoning items, this measure evidenced good reliability ($\alpha = 0.87$) and has demonstrated expected criterion-related validity in past studies (e.g., $r = 0.20$, uncorrected). Scores are average and thus range from 0 to 1.

5.2.3. Blatant extreme responding

BER (Landers et al., 2011) was computed as the proportion of items in a personality measure for which participants endorsed the desirable extreme points of the scale (i.e., 1 or 5). This approach to measure BER as a continuous variable is slightly different from the original binary approach by Landers et al. (2011), which was motivated by a coaching rumor in the studied organization advising people to answer *all* items with 1 and 5 s.[1] BER scores were obtained for each personality measure, and an overall BER score was computed as the average across all personality measures. All BER scores ranged from 0 (when participants did not endorse any desirable extreme points) to 1 (when participants endorsed all desirable extreme points). The average intercorrelation between BER scores for various personality measures was 0.57. This suggests that participants tend to endorse desirable extreme points as a general response style, not only for one personality trait.

5.2.4. Unlikely virtues

The 10-item ($\alpha = 0.92$) UV scale (Weekley, 2006) measured extremely virtuous behaviors (e.g., 'When growing up, I never told my parents a lie'). A 5-point rating scale (*strongly agree* to *strongly disagree*) was used. High scores on this scale suggest that the applicant is responding to the scale with an overly positive self-evaluation.

6. Results

Descriptive statistics and correlations among study variables are reported in Table 1. Analyses of variance (ANOVAs) were used to test Hypothesis 1, suggesting that applicants would score higher than incumbents on personality scales, BER, and UV, but not on GMA (Table 2). Applicants scored higher than incumbents on all personality measures, except *concern for others*. We computed the effect sizes (Cohen's *d*; Cohen & Cohen, 1983) and found that applicants' scores were between 0.53 and 0.99 standard deviation units higher than incumbents on personality measures (i.e., large effect sizes, except again for *concern for others*). Applicants also scored higher than incumbents on the overall BER score ($d = 1.02$), on all BER scores at the personality trait level (d between 0.46 and 1.10, except again for *concern for others*), and on the UV scale ($d = 0.87$). Finally, incumbents and applicants' GMA scores were similar, with slightly higher scores for incumbents ($d = 0.04$). These results provide strong support to Hypotheses 1a, 1b, 1c, and 1d.

In order to test Hypothesis 2, that there would be a positive relationship between overall BER and UV, the zero-order correlation of overall BER with UV was examined. BER showed a significant and medium-sized relationship with UV in both the applicant ($r = 0.39$, $n = 346{,}224$, $p < .001$) and the incumbent ($r = 0.26$, $n = 2{,}295$, $p < .001$) samples. We used Fisher r-to-z transformation to test for significant difference between the two correlation coefficients and found that BER and UV were more strongly related in the applicant than in the incumbent sample ($Z = 6.95$, $p < .001$), providing support to Hypothesis 2.

Hypothesis 3, stating that GMA would correlate significantly higher with personality in the applicant than in the incumbent sample, was supported. Data were available for six personality traits (achievement, initiative, energy, social orientation, persistence, and leadership). For applicants, all correlations with GMA were positive and significant ($rs = 0.05$–0.14, $ns = 10{,}125$–$106{,}231$, $p < .001$). For incumbents, correlations were all nonsignificant ($rs = -0.04$ to 0.00), except for leadership and achievement (both $r = -0.07$, $n = 1{,}606$, $p < .01$). We used Fisher r-to-z transformation to test for difference between the five correlation coefficients and found significantly higher correlations for all five traits in the applicant sample ($Zs = 2.49$–7.12, $ps < .05$).

Hypothesis 4, stating that GMA would positively correlate with BER but would negatively correlate with UV in the applicant sample, was also supported. As expected, the relationship between BER and GMA was positive yet rather small ($r = 0.06$, $n = 106{,}231$, $p < .001$) whereas the relationship between UV and GMA was negative and larger ($r = -0.13$, $n = 102{,}377$, $p < .001$). We used Fisher r-to-z transformation to test for significant difference between two correlation coefficients and found a significant difference ($Z = 42.17$, $p < .001$).

Research Questions 1, 2, and 3 examining job type, gender, and race differences in BER and UV were tested simultaneously with two multiple linear regressions. Participant type (applicant vs. incumbent), job type (managerial vs. nonmanagerial), gender (men vs. women), and race (Black, Hispanic, Asian, and other vs. White) were entered as predictors of BER and UV scores. Because of large differences in samples sizes (e.g., for job type), we used weighted effect coding for our categorical variables (Cohen, Cohen, West, & Aiken, 2003). Main effects were entered in a first step and all (two-way, three-way, and four-way) interactions were entered in later steps. F tests of R^2 change suggested that only two-way interactions significantly

Table 2. Differences between incumbents and applicants for general mental ability, personality, BER, and UV

	Incumbent					Applicant					ANOVA			
Variable	*N*	*M*	*SD*	95% CI		*N*	*M*	*SD*	95% CI		*df*	*F*	*p* value	Cohen's *d*
General mental ability	1,606	0.59	0.23	0.58	0.61	106,231	0.58	0.23	0.58	0.58	1,107835	5.834	0.02	0.04
Achievement	2,295	4.41	0.48	4.39	4.43	65,651	4.67	0.47	4.66	4.67	1, 67944	671.180	<0.001	0.55
Initiative	2,295	4.30	0.52	4.28	4.32	350,078	4.63	0.49	4.63	4.64	1,352371	1066.838	<0.001	0.67
Attention to detail	2,295	4.11	0.58	4.09	4.14	8,120	4.58	0.46	4.57	4.59	1, 10413	1651.382	<0.001	0.90
Energy	2,295	4.01	0.61	3.98	4.03	216,570	4.52	0.52	4.52	4.52	1,218863	2242.320	<0.001	0.90
Concern for others	2,295	4.18	0.61	4.15	4.20	8,120	4.09	0.61	4.08	4.10	1, 10413	37.226	<0.001	0.15
Cooperation	2,295	4.46	0.48	4.44	4.48	235,727	4.75	0.46	4.75	4.76	1,238020	961.532	<0.001	0.62
Self-control	2,295	3.84	0.66	3.82	3.87	214,518	4.43	0.53	4.43	4.43	1,216811	2706.970	<0.001	0.99
Social orientation	2,295	4.18	0.53	4.16	4.20	26,197	4.44	0.45	4.44	4.45	1, 28490	703.120	<0.001	0.53
Persistence	1,853	4.16	0.50	4.14	4.18	80,034	4.50	0.47	4.50	4.51	1, 81885	941.599	<0.001	0.70
Leadership	1,606	4.18	0.60	4.15	4.20	36,322	4.48	0.50	4.48	4.49	1, 37926	580.668	<0.001	0.54
BER overall score	2,295	0.45	0.23	0.43	0.45	350,078	0.70	0.26	0.70	0.70	1,352371	2162.405	<0.001	1.02
BER achievement	2,295	0.60	0.28	0.59	0.61	65,651	0.78	0.24	0.78	0.78	1, 67944	1219.755	<0.001	0.69
BER initiative	2,295	0.47	0.33	0.46	0.48	350,078	0.73	0.30	0.73	0.73	1,352371	1648.356	<0.001	0.82
BER attention to detail	2,295	0.42	0.30	0.41	0.43	8,120	0.73	0.26	0.72	0.73	1, 10413	2283.359	<0.001	1.10
BER energy	2,295	0.40	0.30	0.39	0.41	216,570	0.67	0.31	0.67	0.67	1,218863	1684.456	<0.001	0.89
BER concern for others	2,295	0.48	0.33	0.46	0.49	8,120	0.53	0.31	0.52	0.53	1, 10413	46.719	<0.001	0.16
BER cooperation	2,295	0.58	0.33	0.56	0.59	235,727	0.82	0.26	0.82	0.82	1,238020	1914.995	<0.001	0.81
BER self-control	2,295	0.35	0.29	0.34	0.37	214,518	0.64	0.30	0.64	0.64	1,216811	2058.298	<0.001	0.98
BER social orientation	2,295	0.48	0.29	0.46	0.49	26,197	0.61	0.27	0.61	0.62	1, 28490	538.133	<0.001	0.46
BER persistence	1,853	0.45	0.30	0.43	0.46	80,034	0.66	0.29	0.66	0.66	1, 81885	968.679	<0.001	0.71
BER leadership	1,606	0.45	0.33	0.43	0.46	36,322	0.62	0.32	0.62	0.62	1, 37926	445.820	<0.001	0.52
Unlikely virtue	2,295	2.29	0.86	2.25	2.32	346,224	3.13	1.07	3.13	3.14	1,348517	1430.167	<0.001	0.87

Note: BER = blatant extreme responding; UV = unlikely virtue scale. *BER overall* represents BER across all personality subscales. BER achievement, BER initiative, etc. represent BER for each particular subscale.

Table 3. Base rate for blatant extreme responding across job type, gender, and race

			Incumbent					Applicant				
Job type	Gender	Race	*N*	*M*	*SD*	95% CI		*N*	*M*	*SD*	95% CI	
Nonmanagerial	Men	White	367	0.38	0.22	0.36	0.41	44,192	0.68	0.27	0.68	0.69
		Black	81	0.44	0.22	0.39	0.50	26,617	0.73	0.25	0.72	0.73
		Hispanic	49	0.43	0.25	0.36	0.51	13,970	0.73	0.25	0.73	0.74
		Asian	39	0.47	0.19	0.39	0.55	4,179	0.69	0.27	0.69	0.70
		Other	7	0.36	0.26	0.17	0.54	1,329	0.72	0.25	0.70	0.73
		Total	543	0.40	0.22	0.37	0.46	90,287	0.70	0.26	0.71	0.71
	Women	White	351	0.41	0.21	0.38	0.44	39,620	0.69	0.26	0.69	0.70
		Black	87	0.52	0.20	0.46	0.57	38,696	0.74	0.25	0.74	0.74
		Hispanic	42	0.48	0.22	0.40	0.56	11,566	0.74	0.24	0.73	0.74
		Asian	55	0.44	0.20	0.37	0.51	2,615	0.70	0.26	0.69	0.71
		Other	–	–	–	–	–	1,456	0.70	0.25	0.69	0.72
		Total	535	0.44	0.21	0.43	0.49	93,953	0.72	0.25	0.71	0.72
	Total		1078	0.42	0.22	0.41	0.47	184,240	0.71	0.26	0.71	0.72
Managerial	Men	White	338	0.41	0.24	0.38	0.44	163	0.76	0.21	0.72	0.80
		Black	43	0.47	0.21	0.39	0.54	58	0.81	0.19	0.74	0.88
		Hispanic	43	0.44	0.22	0.37	0.52	36	0.80	0.18	0.72	0.89
		Asian	15	0.52	0.20	0.39	0.65	10	0.79	0.23	0.63	0.95
		Other	–	–	–	–	–	1	0.95	–	–	–
		Total	439	0.42	0.23	0.42	0.50	268	0.78	0.20	0.71	0.93
	Women	White	177	0.44	0.23	0.40	0.48	82	0.75	0.18	0.69	0.80
		Black	30	0.42	0.22	0.33	0.51	27	0.80	0.16	0.70	0.90
		Hispanic	20	0.43	0.24	0.32	0.55	12	0.84	0.11	0.69	0.98
		Asian	14	0.47	0.18	0.33	0.61	4	0.78	0.12	0.53	1.03
		Other	1	0.48	–	–	–	2	0.90	0.14	0.54	1.26
		Total	242	0.44	0.23	0.34	0.56	127	0.77	0.17	0.72	0.91
	Total		681	0.43	0.23	0.39	0.52	395	0.78	0.19	0.75	0.89

contributed to the prediction of BER and UV. Analyses were thus run on these models.

For BER, we found a main effect of participant type ($b = 0.003$, $SE = 0.001$, $p < .01$), but no main effect of race, job type, or gender. We also found a small participant type × job type interaction ($b = 0.001$, $SE = 0.000$, $p < .01$) and a small gender × Black interaction ($b = 0.002$, $SE = 0.001$, $p < .05$). Applicants for managerial positions ($M = 0.78$, $SD = 0.19$) scored higher than applicants for nonmanagerial positions ($M = 0.71$, $SD = 0.26$; $d = 0.31$), but there was no difference for incumbents ($M = 0.43$, $SD = 0.23$ for managerial jobs and $M = 0.42$, $SD = 0.22$ for nonmanagerial jobs; $d = 0.04$). Base rates for all cases can be found in Table 3. Altogether, demographic variables only explained 2% of the variance in BER scores.

For UV, we found main effects of participant type ($b = 0.011$, $SE = 0.004$, $p < .01$), Black ($b = 0.122$, $SE = 0.046$, $p < .01$), Hispanic ($b = 0.058$, $SE = 0.015$, $p < .01$), and Asian ($b = 0.022$, $SE = 0.006$, $p < .01$), but no main effect of job type or gender. We also found small gender × Black ($b = 0.006$, $SE = 0.003$, $p < .05$) and gender × Hispanic ($b = 0.002$, $SE = 0.001$, $p < .05$) interactions. Base rates for all cases can be found in Table 4. White participants ($M = 2.89$, $SD = 1.01$) scored lower than Black ($M = 3.26$, $SD = 1.08$), Hispanic ($M = 3.44$, $SD = 0.98$), or Asian ($M = 3.35$, $SD = 0.94$) participants (d ranging from 0.35 to 0.53), and the difference for Black and Hispanic participants was larger for women. These results suggest race differences for UV scores, but not for BER scores. Altogether, demographic variables explained 6% of the variance in UV scores.

7. Discussion

7.1. Main findings and contribution to personnel selection and faking research

The present research offered a comprehensive examination of BER as a new measure of applicant faking in personality tests. Our field study with a large sample of incumbents and applicants enhances the external validity of our results. Our findings contribute to research on faking on personality measures in several ways.

First, we provide some evidence for the construct validity of BER. As expected, applicants scored significantly higher than incumbents on BER, UV, and personality measures, but not on GMA. These differences were rather large (e.g., $d = 1.02$ for overall BER), suggesting that the motivation to get the job influences test taking behavior and was captured by BER scores. Our results thus allow us to ascertain that BER indeed captures applicant faking, and to eliminate alternative explanations (i.e., true individual differences or systematic ERS) as ac-

Table 4. Base rate for unlikely virtues across job type, gender, and race

Job type	Gender	Race	Incumbent *N*	Incumbent *M*	Incumbent *SD*	Incumbent 95% CI		Applicant *N*	Applicant *M*	Applicant *SD*	Applicant 95% CI	
Nonmanagerial	Men	White	367	2.06	0.79	1.96	2.17	43,769	2.84	1.03	2.83	2.85
		Black	81	2.40	0.90	2.17	2.62	26,506	3.17	1.08	3.16	3.19
		Hispanic	49	2.68	0.79	2.39	2.97	13,923	3.38	1.00	3.36	3.39
		Asian	39	2.72	0.86	2.39	3.04	4,153	3.60	0.98	3.57	3.64
		Other	7	2.27	1.45	1.51	3.03	1,324	3.22	1.05	3.17	3.28
		Total	543	2.22	0.86	2.06	2.38	89,675	3.06	1.06	3.05	3.08
	Women	White	351	2.23	0.81	2.12	2.33	39,387	2.95	0.99	2.94	2.96
		Black	87	2.49	0.87	2.28	2.71	38,573	3.32	1.07	3.31	3.33
		Hispanic	42	2.90	0.93	2.59	3.21	11,537	3.52	0.95	3.51	3.54
		Asian	55	2.84	0.82	2.57	3.11	2,598	3.70	0.93	3.66	3.74
		Other	–	–	–	–	–	1,451	3.25	0.99	3.20	3.30
		Total	535	2.39	0.87	2.23	2.54	93,546	3.20	1.04	3.19	3.21
	Total		1,078	2.30	0.86	2.14	2.46	183,221	3.13	1.06	3.12	3.15
Managerial	Men	White	338	2.00	0.72	1.89	2.11	163	2.58	1.11	2.42	2.74
		Black	43	2.13	0.74	1.82	2.43	58	2.89	1.27	2.62	3.15
		Hispanic	43	2.26	0.82	1.96	2.57	36	3.13	1.13	2.80	3.47
		Asian	15	2.51	0.94	1.99	3.03	10	3.11	0.96	2.47	3.75
		Other	–	–	–	–	–	1	3.90	–	–	–
		Total	439	2.05	0.75	1.89	2.22	268	2.75	1.16	2.51	2.95
	Women	White	177	2.25	0.84	2.10	2.40	82	2.73	1.10	2.51	2.95
		Black	30	2.25	0.95	1.88	2.62	27	3.01	1.31	2.63	3.40
		Hispanic	20	2.63	0.77	2.18	3.08	12	3.25	1.34	2.67	3.83
		Asian	14	2.56	0.96	2.02	3.09	4	3.10	1.32	2.10	4.10
		Other	1	1.80	–	–	–	2	3.75	1.77	2.33	5.17
		Total	242	2.30	0.86	2.07	2.51	127	2.87	1.18	2.53	3.20
	Total		681	2.14	0.80	1.95	2.32	395	2.79	1.17	2.52	3.03

counting for differences in BER because such explanations cannot justify higher scores for applicants than incumbents. Moreover, BER, although an indirect approach to capturing faking, was related to another more direct measure of faking, such as UV. Because faking is an item-level phenomenon (i.e., applicants fake individual items), it should be identified at the item level (Zickar & Robie, 1999), which is precisely what BER allows. In line with past research (Landers et al., 2011; van Hooft & Born, 2012), our results provide further evidence suggesting that extreme responding is a relevant indicator for detecting faking.

Second, BER was (slightly) positively related to GMA whereas UV was negatively related to it. This suggests that applicants with higher GMA are slightly more likely to inflate personality scores (probably because they are better able to identify the desirable answer and use BER as a strategy to increase their chances to get the job). But they are also more likely to deduce the intent of UV items and avoid endorsing them. As such, it seems that high-GMA applicants may not fake more *per se*, but tend to fake more selectively. Moreover, although the relationship between GMA and UV is small, BER could be a more appropriate measure of faking in situations where applicants have higher GMA (e.g., for managerial jobs).

Third, we found that applicants for managerial jobs got slightly higher BER scores than applicants for nonmanagerial jobs, whereas no difference was observed for UV. Applicants for managerial jobs may have had more experience with personality measures than applicants for nonmanagerial jobs, thus having more knowledge of the 1 or 5 s answers to be a good strategy to get higher scores (i.e., similar to the *rumor* effect in Landers et al., 2011).

Finally, we found rather large differences between Whites and non-Whites on UV scores, and this difference was larger for women. These results are similar to Weekley's (2006) findings, although we did not find a main effect for gender. But we found no relevant demographic differences for BER, suggesting that risks associated with adverse impact may be lower for BER than for UV. This may be due to the lower relationship BER has with GMA (as compared to UV). These results also provide further evidence that BER indeed captures faking and not simply systematic ERS, because previous research has showed rather large cultural and racial differences in ERS (that we do not find with our BER scores). For instance, Blacks (Bachman & O'Malley, 1984) and Mexican-Americans (Davis, Resnicow, & Couper, 2011) are more likely than Whites to use extreme responses. East Asians are less likely than Americans to use extreme responses (Chen, Lee, & Stevenson, 1995). Furthermore, masculine or large power distance cultures also tend to use more ERS (Johnson, Kulesa, Llc, Cho, & Shavitt, 2005).

7.2. Practical implications

These results have implications for organizations using self-report personality measures as part of their selection process. One possibility to identify and eliminate fakers would be to identify those applicants who score high on both BER and UV. Removing them from consideration may improve selection decisions and increase the validity of personality scales by taking out of the sample at least some of those faking the personality measures. However, this strategy may have consequences for applicants from minority groups who tend to perform more poorly on measures of GMA (e.g., Chan, Schmitt, DeShon, Clause, & Delbridge, 1997), and thus create adverse impact. Our data suggest that minority participants do score higher on UV, but also confirm that they score lower on GMA than Whites ($d = 0.66$ for Black participants and $d = 0.41$ for Hispanic participants). Using BER alone may thus be a fairer alternative to detect fakers.

Yet, it is still unclear what cutoff (e.g., 80%, 90%, or 100% BER) should be used to identify fakers without eliminating people honestly scoring high on the personality measure. Landers et al. (2011) chose 100% BER (i.e., construing only applicants endorsing all extreme points as fakers). But in practice, lower cutoffs may be more appropriate if organizations want to use a more conservative approach. For instance, in our applicant sample, 17.6% of applicants scored between 80% and 89% on overall BER, 24.5% scored between 90% and 99%, and 7.2% scored 100%. As such, a cutoff of 100% BER would construe only 7.2% of applicants as fakers whereas a cutoff of 80% BER would construe 49.3% of applicants as fakers.

7.3. Limitations and future research directions

The present study does have limitations which should be addressed in future research. Foremost, this study does not include performance measures for applicants. Future studies should include criterion data to examine if eliminating high-BER scorers or using BER- and/or UV-based corrections can help improve the predictive validity of personality measures. Moreover, replication studies could examine the relationship between BER and other measures of faking besides UV, such as SD or other inventory-specific measures of faking (e.g., the *F* scale of MMPI; Hsu, Santelli, & Hsu, 1989), to reinforce confidence in the present findings. Results should also be replicated with other personality measures, such as inventories based on the Big Five (Costa & McCrae, 1992), the Hogan Personality scales (Hogan, 1986), the MMPI (Dahlstrom et al., 1975), or the HEXACO (Lee & Ashton, 2004) models. Moreover, future research should continue to explore the relationship between GMA and BER (or other measures of faking) and uncover the factors leading high-GMA applicants to fake more (because they are more capable to do so) or less (because they think they do not need to do so). One possibility could be to examine the moderating effect of self-esteem.

Faking has been conceptualized as a multidimensional construct, including different types of faking, such as exaggeration, omission, and lying (Levashina & Campion, 2007). As such, it is important to develop different methods to control for different types of faking. For example, bogus items that ask job applicants to assess their familiarity with nonexistent tasks and events (Anderson, Warner, & Spencer, 1984; Pannone, 1984) capture the claim of impossible experiences and thus measure fabrication of information or lies. BER could measure exaggeration if applicants claim that they possess more of a particular trait by endorsing the maximum score on a scale. But it could also capture lying if applicants who actually score extremely low on desirable personality trait chose to endorse extremely positive points to make a better impression. As such, future research may want to determine if/when BER captures exaggeration versus lying, and then if BER corrections can help control for exaggeration or lying, thus complementing other faking detection techniques.

8. Conclusion

Personality measures remain an integral part of many selection systems. Yet, applicants are prone to faking behavior. Therefore, efforts to detect and control for faking are important and should continue. As the current study suggests, BER can be an alternative solution to this vexing problem. It captures differences between applicants and incumbents, is related to a more established measure of faking, and shows less potential for creating adverse impact.

Note

1. We note here that we also reproduced our analyses using Landers et al.'s (2011) approach to compute BER and obtained similar results.

References

Anderson, C. D., Warner, J. L., & Spencer, C. C. (1984). Inflation bias in self-assessment examinations: Implications for valid employee selection. *Journal of Applied Psychology*, 69, 574–580.

Arce-Ferrer, A. J. (2006). An investigation into the factors influencing extreme-response style: Improving meaning of translated and culturally adapted rating scales. *Educational and Psychological Measurement*, 66, 374–392.

Arthur, W., Glaze, R. M., Villado, A. J., & Taylor, J. E. (2010). The magnitude and extent of cheating and response distortion effects on unproctored internet-based tests of cognitive ability and personality. *International Journal of Selection and Assessment*, 18, 1–16.

Austin, E. J., Deary, I. J., & Egan, V. (2006). Individual differences in response scale use: Mixed Rasch modelling of responses to NEO-FFI items. *Personality and Individual Differences*, 40, 1235–1245.

Austin, E. J., Hofer, S. M., Deary, I. J., & Eber, H. W. (2000). Interactions between intelligence and personality: Results from two large samples. *Personality and Individual Differences*, 29, 405–427.

Bachman, J. G., & O'Malley, P. M. (1984). Yea-saying, nay-saying, and going to extremes: Black–White differences in response Styles. *The Public Opinion Quarterly*, 48, 491–509.

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment*, 14, 317–335.

Chan, D., Schmitt, N., DeShon, R. P., Clause, C. S., & Delbridge, K. (1997). Reactions to cognitive ability tests: The relationships between race, test performance, face validity perceptions, and test-taking motivation. *Journal of Applied Psychology*, 82, 300.

Chen, C., Lee, S.-Y., & Stevenson, H. W. (1995). Response style and cross-cultural comparisons of rating scales among East Asian and North American students. *Psychological Science*, 6, 170–175.

Cheung, G. W., & Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. *Journal of Cross-cultural Psychology*, 31, 187–212.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for behavioral sciences* (3rd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Costa, P. T., & McCrae, R. R. (1992). *Revised neo personality inventory (neo pi-r) and neo five-factor inventory (neo-ffi)*. Odessa, FL: Psychological Assessment Resources.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302.

Dahlstrom, W. G., Welsh, G. S., & Dahlstrom, L. E. (1972). *An MMPI handbook, volume 1: A clinical interpretation*. Minneapolis, MN: University of Minnesota Press.

Dahlstrom, W. G., Welsh, G. S., & Dahlstrom, L. E. (1975). *An MMPI handbook, volume 2: Research applications*. Minneapolis, MN: University of Minnesota Press.

Davis, R. E., Resnicow, K., & Couper, M. P. (2011). Survey response styles, acculturation, and culture among a sample of Mexican American adults. *Journal of Cross-Cultural Psychology*, 42, 1219–1236.

De Beuckelaer, A., Weijters, B., & Rutten, A. (2010). Using ad hoc measures for response styles: A cautionary note. *Quality and Quantity*, 44, 761–775.

De Fruyt, F., Aluja, A., García, L. F., Rolland, J.-P., & Jung, S. C. (2006). Positive presentation management and intelligence and the personality differentiation by intelligence hypothesis in job applicants. *International Journal of Selection and Assessment*, 14, 101–112.

de Jong, M. G., Steenkamp, J.-B. E. M., Fox, J.-P., & Baumgartner, H. (2008). Using item response theory to measure extreme response style in marketing research: A global investigation. *Journal of Marketing Research (JMR)*, 45, 104–115.

Donovan, J. J., Dwight, S. A., & Hurtz, G. M. (2003). An assessment of the prevalence, severity, and verifiability of entry-level applicant faking using the randomized response technique. *Human Performance*, 16, 81–106.

Douglas, E. F., McDaniel, M. A., & Snell, A. (1996). *The validity of non-general mental measures decay when applicants fake*. Paper presented at the Annual Conference of the Academy of Management, Cincinnati, OH.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement*, 11, 59–79.

Dudley, N. M., McFarland, L. A., Goodman, S. A., Hunt, S. T., & Sydell, E. J. (2005). Racial differences in socially desirable responding in selection contexts: Magnitude and consequences. *Journal of Personality Assessment*, 85, 50–64.

Dunnette, M. D., McCartney, J., Carlson, H. C., & Kirchner, W. K. (1962). A study of faking behavior on a forced-choice self-description checklist. *Personnel Psychology*, 15, 13–24.

Dwight, S. A., & Donovan, J. J. (2003). Do warnings not to fake reduce faking? *Human Performance*, 16, 1–23.

Egan, V. (1989). Links between personality, ability and attitudes in a low-IQ sample. *Personality and Individual Differences*, 10, 997–1001.

Ellingson, J. E., Sackett, P. R., & Connelly, B. S. (2007). Personality assessment across selection and development contexts: Insights into response distortion. *Journal of Applied Psychology*, 92, 386–395.

Furnham, A. (1986). Response bias, social desirability and dissimulation. *Personality and Individual Differences*, 7, 385.

Furnham, A. (1990). Faking personality questionnaires: Fabricating different profiles for different purposes. *Current Psychology*, 9, 46–55.

Goffin, R. D., & Christiansen, N. D. (2003). Correcting personality tests for faking: A review of popular personality tests and an initial survey of researchers. *International Journal of Selection and Assessment*, 11, 340–344.

Griffith, R. L., & Converse, P. (2011). The rules of evidence and the prevalence of applicant faking. In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.), *New perspectives on faking in personality assessment* (pp. 34–52). New York: Oxford University Press.

Griffith, R. L., & McDaniel, M. (2006). The nature of deception and applicant faking behavior. In R. L. Griffith & M. H. Peterson (Eds.), *A closer examination of applicant faking behavior* (pp. 1–19). Greenwich, NC: Information Age Publishing.

Griffith, R. L., & Peterson, M. H. (2008). The failure of social desirability measures to capture applicant faking behavior. *Industrial and Organizational Psychology*, 1, 308–311.

Guo, J., & Drasgow, F. (2010). Identifying cheating on unproctored internet tests: The Z-test and the likelihood ratio test. *International Journal of Selection and Assessment*, 18, 351–364.

Harold, C. M., McFarland, L. A., & Weekley, J. A. (2006). The validity of verifiable and non-verifiable biodata items: An examination across applicants and incumbents. *International Journal of Selection and Assessment*, 14, 336–346.

Hogan, R. (1986). *Hogan personality inventory manual*. Tulsa, OK: Hogan Assessment Systems.

Holden, R. R., & Hibbs, N. (1995). Incremental validity of response latencies for detecting fakers on a personality test. *Journal of Research in Personality*, 29, 362–372.

Hough, L. M. (1998). Effects of intentional distortion in personality measurement and evaluation of suggested palliatives. *Human Performance*, 11, 209–244.

Hough, L. M., Eaton, N. K., Dunnette, M. D., Kamp, J. D., & McCloy, R. A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology*, 75, 581–595.

Hough, L. M., & Oswald, F. L. (2000). Personnel selection: Looking toward the future-remembering the past. *Annual Review of Psychology*, 51, 631–664.

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment*, 9, 152–194.

Hsu, L. M., Santelli, J., & Hsu, J. R. (1989). Faking detection validity and incremental validity of response latencies to MMPI subtle and obvious items. *Journal of Personality Assessment*, 53, 278–295.

Huffcutt, A. I., Conway, J. M., Roth, P. L., & Klehe, U.-C. (2004). The impact of job complexity and study design on situational and behavior description interview validity. *International Journal of Selection and Assessment*, 12, 262–273.

Hurtz, G. M., & Alliger, G. M. (2002). Influence of coaching on integrity test performance and unlikely virtues scale scores. *Human Performance*, 15, 255–273.

Johnson, T., Kulesa, P., Llc, I., Cho, Y. I., & Shavitt, S. (2005). The relation between culture and response styles: Evidence from 19 countries. *Journal of cross-cultural psychology*, 36, 264–277.

Kuncel, N. R., & Borneman, M. J. (2007). Toward a new method of detecting deliberately faked personality tests: The use of idiosyncratic item responses. *International Journal of Selection and Assessment*, 15, 220–231.

Kuncel, N. R., & Tellegen, A. (2009). A conceptual and empirical reexamination of the measurement of the social desirability of items: Implications for detecting desirable response style and scale development. *Personnel Psychology*, 62, 201–228.

Landers, R. N., Sackett, P. R., & Tuzinski, K. A. (2011). Retesting after initial failure, coaching rumors, and warnings against faking in online personality measures for selection. *Journal of Applied Psychology*, 96, 202–210.

Lao, R. Y. (2001). *Faking in personality measures: Effects on the prediction of job performance*. Dissertation Abstracts International (62).

Lee, K., & Ashton, M. C. (2004). Psychometric properties of the HEXACO Personality Inventory. *Multivariate Behavioral Research*, 39, 329–358.

Levashina, J., & Campion, M. A. (2006). A model of faking likelihood in the employment interview. *International Journal of Selection and Assessment*, 14, 299–316.

Levashina, J., & Campion, M. A. (2007). Measuring faking in the employment interview: Development and validation of an interview faking behavior scale. *Journal of Applied Psychology*, 92, 1638–1656.

Levashina, J., Morgeson, F. P., & Campion, M. A. (2009). They don't do It often, but they do it well: Exploring the relationship between applicant mental abilities and faking. *International Journal of Selection and Assessment*, 17, 271–281.

McFarland, L. A., & Ryan, A. M. (2000). Variance in faking across noncognitive measures. *Journal of Applied Psychology*, 85, 812–821.

McGrath, R. E., Mitchell, M., Kim, B. H., & Hough, L. (2010). Evidence for response bias as a source of error variance in applied assessment. *Psychological Bulletin*, 136, 450–470.

Mersman, J. L., & Shultz, K. S. (1998). Individual differences in the ability to fake on personality measures. *Personality and Individual Differences*, 24, 217–227.

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Are we getting fooled again? Coming to terms with limitations in the use of personality tests for personnel selection. *Personnel Psychology*, 60, 1029–1049.

Möttus, R., Allik, J., Realo, A., Rossier, J., Zecca, G., Ah-Kion, J., et al. (2012). The effect of response style on self-reported conscientiousness across 20 countries. *Personality and Social Psychology Bulletin*, 38, 1423–1436.

O'Connell, M. S., Kung, M.-C., & Tristan, E. (2011). Beyond Impression Management: Evaluating three measures of response distortion and their relationship to job performance. *International Journal of Selection and Assessment*, 19, 340–351.

O'Neill, T. A., Lee, N. M., Radan, J., Law, S. J., Lewis, R. J., & Carswell, J. J. (2013). The impact of 'non-targeted traits' on personality test faking, hiring, and workplace deviance. *Personality and Individual Differences*, 55, 162–168.

Ones, D. S., & Viswesvaran, C. (1998). The effects of social desirability and faking on personality and integrity assessment for personnel selection. *Human Performance*, 11, 245–269.

Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology*, 81, 660–679.

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology*, 78, 679–703.

Pannone, R. D. (1984). Predicting test performance: A content valid approach to screening applicants. *Personnel Psychology*, 37, 507–514.

Pauls, C. A., & Crost, N. W. (2004). Effects of faking on self-deception and impression management scales. *Personality and Individual Differences*, 37, 1137–1151.

Peterson, N. G., Mumford, M. D., Borman, W. C., Jeanneret, P. R., Fleishman, E. A., Levin, K. Y., et al. (2001). Understanding work using the Occupational Information Network (O* NET): Implications for practice and research. *Personnel Psychology*, 54, 451–492.

Ployhart, R. E., & Holtz, B. C. (2008). The diversity-validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology*, 61, 153–172.

Pulakos, E. D., Schmitt, N., Dorsey, D. W., Arad, S., Borman, W. C., & Hedge, J. W. (2002). Predicting adaptive performance: Further tests of a model of adaptability. *Human Performance*, 15, 299–323.

Rosse, J. G., Stecher, M. D., Miller, J. L., & Levin, R. A. (1998). The impact of response distortion on preemployment personality testing and hiring decisions. *Journal of Applied Psychology*, 83, 634–644.

Rothstein, M. G., & Goffin, R. D. (2006). The use of personality measures in personnel selection: What does current research support? *Human Resource Management Review*, 16, 155–180.

Ryan, A., & Sackett, P. (1987). Pre-employment honesty testing: Fakability, reactions of test takers, and company image. *Journal of Business and Psychology*, 1, 248–256.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, 124, 262–274.

Schmitt, N., & Oswald, F. L. (2006). The impact of corrections for faking on the validity of noncognitive measures in selection settings. *Journal of Applied Psychology*, 91, 613.

Snell, A. F., Sydell, E. J., & Lueke, S. B. (1999). Towards a theory of applicant faking: Integrating studies of deception. *Human Resource Management Review*, 9, 219–242.

Tett, R. P., Freund, K. A., Christiansen, N. D., Fox, K. E., & Coaster, J. (2012). Faking on self-report emotional intelligence and personality tests: Effects of faking opportunity, cognitive ability, and job type. *Personality and Individual Differences*, 52, 195–201.

van Hooft, E. A. J., & Born, M. P. (2012). Intentional response distortion on personality tests: Using eye-tracking to understand response processes when faking. *Journal of Applied Psychology*, 97, 301–316.

Viswesvaran, C., & Ones, D. S. (1999). Meta-analysis of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement*, 59, 197–210.

Weekley, J. A. (2006). *Controlling faking in personality measures: Another look at unlikely virtues*. Paper presented at the 22nd Annual Conference of the Society for Industrial and Organizational Psychology, New York, NY.

Weekley, J. A., & McKee, J. (2009). *Controlling faking in personality measures: Yet another look at unlikely virtues*. Paper presented at the 24th Annual Conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Weekley, J. A., Ployhart, R. E., & Cooper-Hakim, A. (2005). *On the development of measures of the O*NET work styles*. Paper presented at the 20th Annual Conference of the Society for Industrial and Organizational Psychology, Los Angeles, CA.

Weekley, J. A., Ployhart, R. E., & Harold, C. M. (2004). Personality and situational judgment tests across applicant and incumbent settings: An examination of validity, measurement, and subgroup differences. *Human Performance*, 17, 433–461.

Wetzel, E., Carstensen, C. H., & Böhnke, J. R. (2013). Consistency of extreme response style and non-extreme response style across traits. *Journal of Research in Personality*, 47, 178–189.

Zickar, M. J., & Drasgow, F. (1996). Detecting faking on a personality instrument using appropriateness measurement. *Applied Psychological Measurement*, 20, 71–87.

Zickar, M. J., & Robie, C. (1999). Modeling faking good on personality items: An item-level analysis. *Journal of Applied Psychology*, 84, 551–563.

Ziegler, M., MacCann, C., & Roberts, R. D. (2011). Faking: Knowns, unknowns, and points of contention. In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.), *New perspectives on faking in personality assessment* (pp. 3–16). New York: Oxford University Press.