Forced-choice pre-employment personality assessment: Construct validity and resistance to faking

Thomas A. O'Neill [a,*], Rhys J. Lewis [b], Stephanie J. Law [a], Nicole Larson [a], Samantha Hancock [c], Jelena Radan [a], Naomi Lee [a], Julie J. Carswell [b]

[a] Department of Psychology, University of Calgary, 2500 University Dr. NW, Calgary, Alberta T2N 1N4, Canada
[b] Sigma Assessment Systems, Canada
[c] Wilfrid Laurier School of Business and Economics, Canada

ABSTRACT

Concerns about faking on pre-employment personality assessment are well founded as respondents often dissimulate. This potentially affects the accuracy of hiring decisions, predictive validity of assessment scores, and defensibility of using personality assessment for high-stakes decision making. We use Jackson's forced-choice instrument and provide novel construct validity evidence. Moreover, our laboratory and field evidence suggests that although respondents are able to elevate test scores in similar magnitudes when instructed, they do not elevate their scores as much when using the forced-choice format.

Crown Copyright © 2016 Published by Elsevier Ltd. All rights reserved.

1. Introduction

Faking on personality tests remains a key barrier to their use in personnel selection. According to Ziegler, MacCann, and Roberts, "Faking represents a response set aimed at providing a portrayal of the self that helps a person to…produce systematic differences in test scores that are not due to the attribute of interest" (2011, p.8). Faking on personality tests for employee screening ultimately affects the validity of the measures and their ability to predict future job performance as well as affect the quality of hiring decisions.

There are three primary strategies used to investigate personality test faking. First, mean sample-level trait scores of job applicants are compared to individuals instructed to respond honestly. Second, concurrent and criterion correlations are compared across applicant and honest response conditions. Third, social desirability scales are included in assessments and means across conditions or partial correlations are investigated. Although the criterion validity of personality holds up under conditions that would encourage faking (e.g., Ones, Viswesvaran, & Reiss, 1996), mean levels on social desirability scales are higher under faking instructions (e.g., Paulhus, Bruce, & Trapnell, 1995). Current findings indicate that individuals can distort their personality test responses, suggesting that the accuracy of hiring decisions informed by personality assessment could be improved.

Rothstein and Goffin (2006) identified the forced-choice (FC) format as a promising option for addressing faking. The FC format presents statements in item sets where respondents identify a subset of statements that are most (least) consistent with their personality. Statements are equated on desirability with the goal of inhibiting faking ability. Whereas some evidence suggests FC can reduce faking (e.g., Christiansen, Burns, & Montgomery, 2005), other evidence shows variability across traits and situations in faking reduction (e.g., Heggestad, Morrison, Reeve, & McCloy, 2006).

The current research focuses on a FC personality assessment developed specifically for pre-employment testing. In Study 1 we offer test–retest reliability and convergent validity evidence. In Study 2 we report on concurrent correlations with measures of social desirability, general mental ability (GMA), and dispositional intelligence (Christiansen, Wolcott-Burnam, Janovics, Burns & Quirk, 2005) considering simulated and actual job applicants. Importantly, we tease apart ability and motivational factors to investigate which factors affect faking behavior in actual job applicants.

2. Study 1

Jackson, a pioneer of personality test development (e.g., Jackson, 1967, 1971, 2000), created the FC measure referred to as the *Employee*

* Corresponding author.
 E-mail address: toneill@ucalgary.ca (T.A. O'Neill).

Screening Questionnaire (ESQ). The ESQ was intended to minimize response distortion when respondents may be motivated to deceive (Jackson, Wroblewski et al., 2000). Items were equated on social desirability and limited to employment-relevant behaviors. The ESQ predicted counterproductive workplace behavior, job satisfaction, and turnover (Iliescu, Ilie, & Ispas, 2011), supporting predictive validity.

A challenge of FC tests is that coefficients of internal consistency assume independent measurement of scales that is not present on forced choice assessments (Hicks, 1970). Consequently, internal consistency coefficients become inappropriate to interpret, although new research in item response theory has been applied to pairwise FC comparisons (e.g., Chernyshenko et al., 2009). In the case of tetrad comparisons (i.e., the ESQ), test–retest reliability is an alternative because it estimates stability of scores over time (Heggestad et al., 2006). A second challenge with FC measures involves their convergent validity with traditional formats, as ipsativity raises questions about meaningful between-person comparisons. However, high correlations with standard measures of the same personality traits would suggest construct similarity and meaningful between-person comparisons (Bartram, 1996).

To investigate the ESQ's convergent validity, we considered two traditional personality assessments (the Six-Factor Personality Questionnaire [SFPQ; Jackson, Paunonen & Tremblay, 2000] and the International Personality Item Pool [IPIP; Goldberg, 1999]) in which items are responded to sequentially (i.e., single-stimulus; SS). We considered the traits Agreeableness (generally defined as sympathetic, kind, trusting, and warm) Extraversion (generally defined as talkative, assertive, outgoing, and sociable), Industriousness (generally defined as ambitious, determined, energetic, and aspiring), and Methodicalness (generally defined as prudent, organized, deliberative, and meticulous) as they are desirable for most jobs.

2.1. Method

Participants were 121 students enrolled in undergraduate psychology courses in a large North American university. Participation involved two sessions. During Session 1 respondents completed the IPIP (10 items per trait) then the ESQ (27 item quartets); in Session 2 (scheduled three weeks later) they completed the SFPQ (16 items per trait) then the ESQ. Fig. 1 contains an ESQ sample item. Respondents endorse an item as most characteristic of themselves (scored $+1$) and least characteristic of themselves (scored -1; remaining items scored 0). The SFPQ

and IPIP employed a five-point Likert-type scale, ranging from 1 (*strongly disagree*) to 5 (*strongly agree*).

3. Results

Test–retest reliability coefficients of the ESQ ranged from .70 to .86. Table 1 contains convergent correlations ranging from .49 to .83. The one exception involved ESQ and IPIP Agreeableness ($r = .10$). We examined the content of the scale items and IPIP Agreeableness assesses empathy, social interest, and social grace, whereas ESQ and SFPQ Agreeableness items relate to even-tempered, good-natured, and accommodating tendencies. Because SFPQ Agreeableness demonstrated strong convergent validity ($r = .71$), we felt that the ESQ Agreeableness scale was functioning well. Accordingly, this adds to existing psychometric support for the ESQ and for pursuing it in a broader pre-employment personality testing study.

4. Study 2

4.1. Score inflation

Studies investigating score inflation on FC test formats in applicant samples are considerably few. Jackson, Wroblewski et al. (2000) reported that the SS format was associated with a three-fold inflation of personality scores compared to the ESQ in a simulated applicant study. Similarly, Christiansen, Burns et al. (2005) found lower FC relative to SS test means for simulated applicants compared to individuals instructed to respond honestly. Heggestad et al. (2006), however, found mixed support for FC versus SS across traits. In one field study using military samples, Stark et al. (2014) reported that a FC test yielded similar score magnitudes when applicants were informed it was for selection versus research purposes. Accordingly, there appears to be potential in the use of FC for controlling score inflation, but there are too few studies to deliver a strong verdict.

4.2. Concurrent validities

Jackson, Wroblewski et al.'s (2000) research compared traits measured with FC and SS formats in the prediction of workplace delinquency. With the SS format, the criterion correlation for trait dependability was halved in the simulated applicant vs. respond honestly condition.

For each set of statements, please indicate which statement MOST describes you, and which statement LEAST describes you.

Fig. 1. Sample ESQ item.