Symposium
TITLE
Practical Considerations for Implementing Personality Testing in Organizations

ABSTRACT
Great care must be taken by organizations as they develop policies and procedures governing selection systems. While substantial research demonstrates the value of personality testing in organizations, less research addresses practical issues associated with the implementation of personality testing. The research presented in this symposium seeks to inform implementation decisions.

PRESS PARAGRAPH
Many companies use personality tests as part of the process of deciding whom to hire. Considerable research has shown that such tests provide organizations with information relevant to making these decisions. However, substantially less research has been conducted on issues related to the ways to implement these tests. Such research is important because implementation procedures can dramatically influence the extent to which applicant’s personality test scores are useful to the organization. In this symposium, researchers will present data that can help organizations design policies and procedures to improve the quality of their testing programs.
Personality testing in organizational research and HR practice has had a long and tumultuous history. After a lengthy period of excitement and widespread use beginning in the early 1900s, a failure to identify clear patterns of validity for personality measures in the prediction of job performance (Ghiselli & Barthol, 1953; Guion & Gottier, 1965) resulted in a virtual moratorium on personality testing in the mid 1960s. Personality testing was given new life in the mid 1980s, fueled by the emergence of the five-factor taxonomy of personality and subsequent meta-analytic findings of moderate predictive validities for these broad personality dimensions (Barrick & Mount, 1991; Ones, Viswesvaran, & Schmidt, 1993; Tett, Jackson, & Rothstein, 1991). As noted by Hough and Oswald (2000), there is currently great optimism for personality assessment in personnel selection. Indeed this optimism has led many organizations to incorporate personality tests into their selection systems.

Given increasing legal concerns and the high-stakes associated with selection decisions, organizations must take great care to develop and implement informed policies and procedures regarding their selection systems. While considerable research has examined the relations between personality traits and a variety of organizational outcomes, less research has been directed toward understanding practical issues associated with the implementation of personality testing policies and procedures within a selection context. To this end, the present symposium is to bring together applied and basic research capable of informing organizations as they implement personality testing into their selection systems.
The Reynolds and Sinar paper addresses issues associated with providing applicants with test information, or coaching, prior to the administration of a non-cognitive test battery. In addition to looking at how coaching impacts applicant scores, their unique applied samples allow for the investigation of the impact of coaching on a variety of items formats (e.g., personality, biodata, situational judgment). Perhaps the most significant finding in this research is the evaluation of how coaching technique and test content interact with subgroup differences. The results of this research have important implications for organizations as they consider whether or not to provide information to applicants regarding the nature of the personality assessments included in their selection battery.

The Ellingson and Heggestad paper explores a method for handling candidates who have been flagged for having responded in an overly socially desirable manner. That is, while many personality tests include measures of social desirability, and the associated manuals include “cutoff scores” for the social desirability scales that indicate which scores are aberrantly high, few manuals offer suggestions for how to handle these flagged applicants. One often-discussed option has been to inform the applicant that he or she has been flagged and to retest the individual. The paper presented by Ellingson and Heggestad provides an empirical evaluation of this recommendation. The results of this study will inform organizations about the viability and consequences of implementing retesting policies for individuals with aberrant social desirability scores.

Many organizations are concerned about the possibility that intentional response distortion will result in a loss of validity for personality tests that have previously been found to be valid in incumbent samples. Arguing that they are less susceptible to the effects of faking, some researchers have advocated the use of forced-choice response formats instead of more
traditional Likert-type response formats. While forced-choice response formats have been shown to be resistant to the effects of faking, concerns remain about their capability to provide assessments of normative trait standing. The paper by Morrison, Heggestad, Reeve and McCloy provides an empirical evaluation of the extent to which a particular forced-choice measure is capable of capturing normative trait standing. The results of this research can be used by organizations as they contemplate whether to employ a forced-choice response format to combat the possibility of intentional response distortion among applicants.

Finally, the paper by Sydell and Snell explores the extent to which perceptions held by the test taker can influence the validity of a personality assessment. In this research, over 300 employees in three different manufacturing organizations participated in a criterion-related validity study and completed a questionnaire assessing a variety of perception and demographic variables. Results indicated that some of these perception and demographic variables significantly moderated the validity of the personality assessment. These findings may be particularly important for organizations as they evaluate whether observed validities for a particular assessment will hold in their applicant pool.

The discussant, D. Brent Smith, will provide his evaluation of the research presented and make general recommendations for organizations considering the incorporation of personality tests into their selection systems.

References


The operational use of non-cognitive measures in personnel selection raises a variety of issues associated with the potential for score manipulation. Often, this issue is examined under the paradigm of faking, because it is presumed that most applicants will distort their ratings to improve their chances of success in the selection program. Less frequently examined (and potentially more extreme) is the influence of formal coaching and orientation programs on non-cognitive measures; however, as the use of these measures regains momentum in applied contexts, the proliferation of coaching programs for these measures will accelerate. Our research focuses on the impact of these programs on various operational outcomes for a range of non-cognitive measures.

Test preparation and orientation programs, while frequently used in certain areas of high-stakes testing, such as standardized aptitude tests and proficiency exams (Sackett, Schmitt, Ellingson, & Kabin, 2001), are less common in employment settings. Job candidates often receive no information prior to their arrival at the testing site, and seldom receive detailed information about the tests they are about to take. This trend has persisted despite findings from applicant reactions research highlighting the potential procedural justice, corporate image, and utility benefits of distributing full and accurate information about an employment testing process (Gilliland & Cherry, 2000). The employment context provides a rich, yet under-explored, test-bed for investigations of the impact of test preparation because they are more likely to include a wide-range of measurement targets and methods; other contexts tend to be more limited to achievement and knowledge measurement.
There are good reasons to suspect that test preparation and orientation programs may have different effects for different types of measurement. The coaching literature has consistently found that subjects can alter their scores on a personality measure after simple instructions to do so (Hough et al., 1990). More recent research suggests that personality and other non-cognitive selection tools vary in the extent to which their scores can be influenced by pre-test instructions or coaching (e.g., McFarland & Ryan, 2000).

Test orientation programs may have real benefits: not only can they enhance perceptions of procedural justice among candidates (Gilliland & Cherry, 2000), they may also mitigate score differences between members of minority and majority groups. In personnel selection research studies, results regarding subgroup differences have been mixed, with some studies finding reductions (e.g., Schmit, 1994), while others have found evidence for increased subgroup differences (Ryer, Schmidt, & Schmitt, 1999). These studies have focused on cognitively oriented measures, however. Research on the impact of test orientation programs using a wider range of test content (e.g., personality scales, situational judgment tests) administered to the same individuals can inform our understanding of how these programs could help to reduce subgroup differences and how the effect may vary by type of measure.

This study compares results from operational testing programs to examine how different types of employment tests orientation programs will (a) have stronger or weaker effects on test score levels, (b) have differential effects on test intercorrelations and construct validity, and (c) have more or less impact in reducing subgroup differences. The study design provides a range of orientation conditions and explores their effects on differences among different demographic groups. In this study, test content (situational judgment, personality, biographical data, conditional reasoning) remains constant across conditions, but the specific form of orientation
varies substantially. These conditions include: (1) an interactive session in which candidates complete a detailed test training program, including personalized feedback from an instructor on areas in which test performance can be improved (n = 1,311), (2) a written orientation packet distributed to participants and containing basic information about the test and the testing process, including sample items of a similar type to that found on the actual employment test (n = 3,247), and (3) no orientation (n = 1,923). All participants completed an identical set of items during the application process for manufacturing positions with several similar organizations.

The central research questions explored in this study include: (1) what are the overall effects of various orientation conditions on candidate scores? (2) How do these orientation effects vary by test type? (3) What are the effects of orientation conditions on subgroup differences? And, (4) what are the interactive effects between these variables? Findings will be discussed in light of their implications for the operational use of each content type under various coaching conditions; for example, at what point does a coaching program produce unintended detrimental effects that outweigh the perceptual benefits associated with procedural justice?

References


Social desirability research gains importance when it addresses practical concerns faced by organizations that use personality inventories to enhance their decision-making process. As self-report measures, personality inventories have the potential to be influenced by socially desirable responding. Most organizations view this potential as an unavoidable detriment, a perspective that significantly reduces the instrumentality of personality measures as decision-making tools. Yet, there is growing empirical evidence that personality measures are valuable predictors of individual job performance (e.g., Hough, 1992; Mount & Barrick, 1995; Salgado, 1997). The implications of this evidence for bridging research and practice suggest that organizations should use personality measures when making hiring decisions. Given organizations’ fears about faking, however, any such recommendation must be paired with suggestions regarding how to manage the issue of socially desirable responding. This presentation will target a key question faced by practitioners: Is the retesting of individuals flagged as socially desirable respondents a viable approach for controlling faking tendencies?

Social desirability research has posited a number of alternatives for ameliorating the impact of faking, such as warnings to respondents that there are consequences associated with distorting responses and that such responses can be detected, the construction of forced-choice or empirically-keyed instruments containing more subtle items, and the application of social desirability corrections (e.g., Christiansen, Goffin, Johnston, & Rothstein, 1994; Hough, 1998; Paulhus, 1981). However, research on these approaches has been inconclusive, suggesting that
individuals may continue to respond in a socially desirable manner even if organizations choose to adopt such approaches.

Another alternative for controlling the impact of faking on organizational decision-making is retesting. Many personality measures designed for organizational use include a social desirability scale for administration in conjunction with traditional personality questions. Test users are often instructed by the manuals for these personality measure to use this scale as a means of tracking and red-flagging those individuals who may have faked their responses (e.g., Conn & Rieke, 1994; Gough & Bradley, 1996; Schmit, Kihm, & Robie, 2000). Yet, these manuals typically offer little advice regarding what to do with red-flagged individuals, leaving organizations faced with the knowledge that an individual has likely distorted his or her responses, but without clear direction as to a reasonable course of action. When organizations encounter red-flagged individuals, choosing to indicate to those individuals that they have been identified as responding in a socially desirable manner, and allowing them to retake the test may serve to reduce the extent of impression management on the second administration. Assuming that response patterns are altered accordingly, decisions concerning the individuals could be made using a more honest set of profile scores.

Little research has gone beyond simple reliability measurement to document the implications of having individuals retake a personality measure. This presentation will outline a current research project that investigates how red-flagged individuals’ scale score profiles change when asked to complete a personality measure again. In addition, since many organizations use warnings to emphasize the importance of responding honestly and the potential for detection should one fail to do so, this research also looks at the impact of retesting when it is done in conjunction with the administration of warnings. As a means of measuring potential
applicant reactions, participants’ attitudes toward retesting in terms of fairness perceptions and job choice are also assessed. Currently, data are being obtained from a sample of roughly 500 undergraduate students. The participants are being subjected to a repeated measures design wherein they twice complete a personality test and candidness scale under motivating instructions. Based on candidness scale scores, flagged participants are retested with instructions to complete the personality measure again, more honestly. Non-flagged participants are also retested as a baseline for observing changes between first administration and second administration scale scores. All participants are asked to complete a survey of their attitudes toward retesting. As an overlay to this design, half of the participants complete the two assessments under warning conditions, while the other half complete the assessments without receiving a warning.

Preliminary analyses of the data will include the computation of descriptive statistics, effect sizes, and $t$-tests to assess the extent and significance of changes in score profiles across administrations and between the two conditions. Further, the percentage of flagged participants will be compared across administrations and conditions, and participants’ scale scores from both administrations will be subjected to mock top-down selection rankings. Finally, respondent reactions will be evaluated. Based on the analyses, a series of results and discussion points will be provided.

References


Do Multidimensional Forced-Choice Measures Provide Normative Data?

Morgan Morrison & Eric D. Heggestad
Colorado State University

Charlie L. Reeve
Purdue University

Rodney A. McCloy
Human Resources Research Organization

Moderate criterion-related validities (e.g., Barrick & Mount, 1991) observed for personality assessments make them attractive for inclusion in personnel selection systems. Persistent concerns about the susceptibility to response distortion, however, have tempered that attractiveness (Graham, McDaniel, Douglas, & Snell, 2002; Mueller-Hanson, Heggestad, & Thornton, in press). Numerous strategies, such as the inclusion of warning statements or statistical corrections based on social desirability scales, have been proposed to mitigate the potential negative effects of faking. Unfortunately, most of these strategies have met with mixed success (Ellingson, Sackett, & Hough, 1999; Haymaker & Erwin, 1980; Lautenschlager & Atwater, 1986; Schrader & Osburn, 1977). Recently, there has been increasing optimism that multidimensional forced-choice (MFC) response formats may reduce applicants’ capability to respond in an overly socially desirable manner (Christiansen, Edelstein, & Fleming, 1998; Jackson, Wrobeleski & Ashton, 2000; White & Young, 1998). Jackson et al. (2000), for instance, found that their MFC measure showed less score inflation and greater criterion-related validity in an applicant context then a comparable Likert scale.

MFC items generally consist of two or more statements, each reflecting a different trait to which the respondent must choose, rank order, or otherwise indicate preference among the statements. Two characteristics of MFC items make them resistant to the effects of faking. First,
the general format of the item prevents the respondent from rating himself or herself highly on each construct (as indicated by the statements) included in the item. Second, MFC items can be created such that each statement within the item is roughly equivalent in terms of social desirability. With no one statement any more desirable than any other, the possibility of responding based on social desirability is considerably reduced.

Although research has shown that MFC measures do reduce the potential for faking, the optimism for these measures must be tempered by potentially serious measurement limitations. Specifically, the selection context necessarily demands assessment of inter-individual differences, which is provided by normative data (such as that provided by Likert-type responses). Unfortunately, because of the dependencies built into the response to any single MFC item, MFC measures provide ipsative or partially ipsative data. While ipsative data are appropriate for intra-individual differences assessment and not inter-individual differences assessment, the extent to which partially ipsative measure provide data appropriate for inter-individual differences assessment has not been fully addressed empirically. Our ongoing research is designed to examine this question directly. Specifically, our research compares the measurement precision of the same set of items administered in a traditional Likert format and the MFC format in both honest and faking conditions to determine if MFC items provided data appropriate for selection.

Currently, data are being collected on approximately 600 individuals who are asked to complete a set of items from the International Personality Item Pool (IPIP; Goldberg, 1999) in Likert and MFC formats (i.e., the response formats differ but the statements representing the traits are identical), as well as the NEO Five Factor Inventory (NEO-FFI; Costa & McCrae 1992). The MFC measure employs the dichotomous quartet method (Dunnette, 1962; Jackson et
al., 2000). Half of the sample is completing the three measures honestly, whereas the other half is completing the NEO honestly but is being instructed to complete the IPIP Likert and MFC measures as if they were applying for a job.

Three sets of analyses are to be performed. First, effect sizes between honest and faking conditions will be calculated and compared for the IPIP Likert and MFC measures to determine whether the MFC measure is indeed resistant to the effects of faking. Second, correlations between comparable scales of the NEO and the IPIP Likert and MFC measures will be evaluated in both the honest and faking conditions. If the MFC measure is adequately capturing normative information, the NEO-MFC correlations should be as strong as the NEO-IPIP Likert correlations. Third, decision consistency analysis (i.e., the degree to which an individual passes/fails two or more administrations of a test will be conducted. Specifically, we will set three hypothetical cut scores, corresponding to the 20th, 50th and 80th percentiles of each measure, and evaluate the consistency of pass/fail decisions across the forms.

The results of this research will be useful for practitioners concerned about applicant response distortion on personality measures using traditional Likert-type response formats. Based on our findings, initial conclusions can be drawn regarding the measurement properties of MFC measures. In addition, recommendations can be provided to practitioners considering the inclusion of MFC measures in their personnel selection systems.

References


For decades, researchers have debated the nature of personality. At one end of the debate is the belief that personality can be represented by relatively stable and meaningful traits (e.g., Goldberg, 1993). Opposing this view is the belief that traits do not exist at all and that personalities are unstable and difficult to define (e.g., Kroger & Wood, 1993). Traditionally, tests of personality used in selection contexts have been assumed to provide indications of the applicant’s true scores on the constructs assessed. Recent writings, however, have suggested that this perspective is naive, and have identified a host of complications (e.g., Mischel & Shoda, 1995). These complications include the realization that individuals do not respond to a set of items in the same way in different situations, and that the concept of true scores may not even exist in personality assessment. Additionally, selection researchers have found that characteristics of the situation can influence test scores. For example, Schmit and Ryan (1992) reported that higher test-taking motivation was related to a decrease in criterion-related validity.

Applied psychologists involved in validity study data analyses routinely face situations in which they are unable to fully validate predictor measures as well as develop a complete understanding of the relationships between predictors and criteria in the sample organization. The greater the understanding of the organizational context, the better chance the psychologist has to discover a valid and meaningful relationship between predictor and criterion. Where
criterion-related validity was once thought to be stable across sample types (e.g., concurrent versus predictive), research has highlighted many factors that hold the potential to influence validity study variable relationships (e.g., Mischel & Shoda, 1995; Snell et al., 1999).

The present study was designed to shed further light on this topic by defining and measuring a variety of personal and perceptual factors that might affect criterion-related validities of personality assessments. The hypotheses stem from a Connectionist-inspired model of personality, which recognizes that responses to stimuli, such as tests, are generated from a nearly unlimited number of transitory, situation-specific influences. The validity of a personality test is thus expected to be influenced by subtle characteristics of the situation and applicant perceptions.

Data were collected as part of concurrent criterion-related validity studies conducted by a human resources consulting firm. The present study utilizes samples from three separate manufacturing organizations; with a total sample size was 305 individuals. In addition to completing the personality measure, incumbents responded to a demographics questionnaire and a reaction form, which measured the perceptual factors examined as potential moderators. Hypotheses were generated around the following objective characteristics of the respondent: experience on the job, gender and age. Hypotheses were also written around the following more subjective factors: knowledge of the job; knowledge of the construct; perceptions of others’ behaviors; perceptions of others’ attitudes; perceptions of fairness/justice; attitudes and behaviors dealing with faking; perceptions of selection system acceptability; perceived expectancy for success; and perceived importance of the outcome.

The data were analyzed in two phases. First, zero-order correlations were examined between the above factors and personality scores. Second, hierarchical moderated regressions
were conducted to understand whether the factors moderate the validity of the personality scales for predicting job performance. The results showed support for several hypotheses and interesting reversals on others. In general, there was more support for hypotheses dealing with the objective factors, and less support for hypotheses dealing with the subjective factors. The findings are discussed and several potential explanations are offered. For example, the results serve to shed new light on the complexities involved in validating non-cognitive measures, and suggest useful future research steps to help further delineate relationships that are usually complex and sometimes counterintuitive. Finally, several important implications from these results can be drawn for practitioners who are engaged in concurrent validity studies.

References


Participant List

Eric D. Heggestad  
Assistant Professor  
Colorado State University  
Department of Psychology  
Colorado State University  
Fort Collins, CO 89523-1876  
970-491-4940  
heggesta@lamar.colostate.edu  
Chair, Co-author

Douglas H. Reynolds  
Manager, Assessment Technology  
Development Dimensions International  
1125 Washington Pike  
Bridgeville, PA 15017  
412-220-2845  
doug.reynolds@ddiworld.com  
Presenter

Evan F. Sinar  
Consultant, Assessment Technology Group  
Development Dimensions International  
1125 Washington Pike  
Bridgeville, PA 15017  
412-220-7207  
evan.sinar@ddiworld.com  
Co-author

Douglas E. Haaland  
Consultant  
Development Dimensions International  
33478 Harper Ave.  
Clinton Township, MI 48035  
586-285-0773  
doug.haaland@ddiworld.com  
Co-author

Jill E. Ellingson  
Assistant Professor  
Department of Management and Human Resources  
The Ohio State University  
734 Fisher Hall  
2100 Neil Avenue
Columbus, OH 43210
614-292-4585
ellingson@cob.osu.edu
Presenter

Morgan Morrison
Graduate Student
Colorado State University
Department of Psychology
Colorado State University
Fort Collins, CO 89523-1876
970-491-4325
morganm@holly.colostate.edu
Presenter

Charlie L. Reeve
Assistant Professor
Department of Psychological Sciences
Purdue University
West Lafayette, IN 47907
765-496-2141
creeve@psych.purdue.edu
Co-author

Rodney A. McCloy
Principal Staff Scientist
Human Resources Research Organization (HumRRO)
66 Canal Center Plaza, Suite 400
Alexandria, VA 22314-1591
703-706-5653
rmccloy@humrro.org
Co-author

Eric J. Sydell
Principal
Shaker Consulting Group, Inc.
26000 Cannon Road
Cleveland, OH 44146
440.439.3500
eric.sydell@shakercg.com
Presenter

Andrea F. Snell
Assistant Professor
The University of Akron
Arts & Sciences Building
D. Brent Smith  
Assistant Professor of Management and Psychology  
Jesse H. Jones Graduate School of Management  
Rice University  
P.O. Box 2932 MS-531  
6100 S. Main Street  
Houston, TX 77252-2932  
713-348-5386  
smithb@rice.edu  
Discussant