*Industrial and Organizational Psychology*, **1** (2008), 291–295.


COMMENTARIES

# Lessons Learned in Transitioning Personality Measures From Research to Operational Settings

LEONARD A. WHITE AND MARK C. YOUNG
*U.S. Army Research Institute for the Behavioral and Social Sciences*

ARWEN E. HUNTER
*The George Washington University*

MICHAEL G. RUMSEY
*U.S. Army Research Institute for the Behavioral and Social Sciences*

Hough and Oswald have acknowledged the major contribution of the U.S. Army's Project A to our understanding of personnel selection within the field of industrial–organizational psychology. Results from validation of the Assessment of Background and Life Experiences (ABLE) developed in Project A provided strong evidence of the utility of personality constructs for predicting important aspects of military performance. Since Project A, the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) has conducted much research on the use of personality measures for personnel selection and classification decisions.

Correspondence concerning this article should be addressed to Leonard A. White. E-mail: len.white@cox.net

Address: U.S. Army Research Institute for the Behavioral and Social Sciences, 2511 Jefferson Davis Highway, Arlington, VA 22202-3926.

Leonard A. White and Mark C. Young, U.S. Army Research Institute for the Behavioral and Social Sciences; Arwen E. Hunter, The George Washington University; Michael G. Rumsey, U.S. Army Research Institute for the Behavioral and Social Sciences.

All statements expressed in this commentary are those of the authors and do not necessarily reflect the official opinions of the U.S. Army Research Institute or the Department of the Army.

Our commentary will focus primarily on the use of personality measures for high-stakes, Army applicant screening. Much of this research has involved either the ABLE or the Assessment of Individual Motivation (AIM), both self-report personality measures assessing constructs that overlap with the Big Five. With respect to item format and transparency, ABLE is similar to many personality measures that are widely used today. AIM uses a forced-choice approach to help reduce concerns regarding fakability.

Without question, the "faking problem" has been one of the greatest challenges to the Army's ability to implement and sustain the operational, large-scale use of self-report personality measures, especially in high-stakes testing situations. Our focus on this issue began during Project A when promising findings resulted in the ABLE being seriously considered for use in high-stakes, Army applicant screening. However, due primarily to concerns about its susceptibility to faking and coaching, ABLE was never used operationally by the Army for applicant screening. Today—20 years later—the Army is having some success in using personality measures for making real-life personnel selection and assignment decisions. In the spirit of Hough



https://doi.org/10.1111/j.1754-9434.2008.00049.x Published online by Cambridge University Press

and Oswald's invitation to exchange ideas, we briefly share our key findings from these decades of research, with the goal of suggesting areas to advance future research.

## Summary of Key Findings

1. High levels of faking, detected by social desirability scales, can lead to highly inflated test scores that have little or no criterion-related validity. In high-stakes testing, where faking is likely to be a concern, new fake-resistant measures are needed along with new approaches to counter faking.

Initial findings from the Project A concurrent validation sample indicated that the criterion-related validity of personality constructs remained stable even when socially desirable responding was elevated (Hough, Eaton, Dunnette, Kamp, & McCloy, 1990). However, analyses of the much larger Project A longitudinal validation research sample showed a very different pattern of findings, with high levels of socially desirable responding severely attenuating the predictive validity of personality constructs across multiple criteria (White, Hunter, & Young, 2006; White, Young, & Rumsey, 2001).

When faking was not high, uncorrected validities ranging from .15 to .30 were obtained for predicting job effort, leadership, personal discipline (i.e., staying out of trouble), and maintaining physical fitness. Personality measures conceptually related to the criterion had the highest validity. For example, work orientation was the best predictor of job effort and non-delinquency had the highest correlation with personal discipline (White et al., 2001). However, when faking was high, the validities of these measures were lower and typically close to zero. In these relationships, social desirability operated as a moderator variable, interacting with the personality predictor scales, with the magnitude of the corresponding validity coefficient varying by degree of socially desirable responding.

Research following Project A (White & Kilcullen, 1998) showed similar moderator effects in four independent military and civilian samples. In this research, the uncorrected criterion-related validity of personality constructs averaged $r = .28$ for individuals showing low and moderate levels of socially desirable responding but fell to $r = -.09$ for individuals with high levels of socially desirable responding.

In these and other research samples, the percentage of cases showing elevated levels of socially desirable responding is often low, but in our experience, the amount of such responding can increase substantially in high-stakes settings, with an associated loss of validity (Young, White, Heggestad, & Barnes, 2004). Faking that results in inflated test scores makes it more likely that an applicant will pass any feasible selection standard chosen for the job. Even small percentages of high fakers are of concern when screening high-risk applicants or selecting elite personnel where the costs of improper selection decisions may be high.

Some investigators have argued that socially desirable responding may be indicative of social competence or measure valid trait variance in Conscientiousness or adjustment and thereby be positively related to subsequent job performance. In our large samples, we have found little support for the notion that individuals who distort their responses perform better on the job or are more adaptable than those who do not (White et al., 2001).

2. It is difficult to simulate the pressures of high-stakes testing, and results of faking experiments can underestimate the score inflation subsequently observed in such settings.

To help reduce concerns about ABLE's fakability and coachability, the Army developed the AIM. AIM is a forced-choice instrument that measures the job-relevant personality constructs from ABLE. Results from several simulated applicant faking experiments to evaluate AIM's fakability were encouraging. In these experiments, when our standard operational warning statements were used, we found only a .1 *SD*

https://doi.org/10.1111/j.1754-9434.2008.00049.x Published online by Cambridge University Press

increase in AIM scores relative to research norms, which also compared favorably against other self-report measures of personality. Later, when AIM was subsequently implemented for operational screening, we observed a large rise in test scores (about 1 *SD*), relative to research norms (Young et al., 2004). The resistance to faking indicated by our simulated applicant faking experiments simply did not materialize in the operational environment.

3. In the personality domain, be wary of generalizing results across contextual boundaries.

There are a variety of contextual boundaries limiting the generalizability of personality measure findings that one ignores at one's peril. The first boundary separates divergent types of research designs: specifically, concurrent versus predictive. We first encountered this issue in the Army's Project A, where the concurrent validities were not always consistent indicators of the magnitude of the predictive relationships. For some items and their associated constructs, there was a close correspondence between the predictive and the concurrent validities, but for several others, the predictive validities were nearly 50% lower (White & Moss, 1995). This finding clearly pointed to a need for more longitudinal validation research in the personality domain.

The second boundary separates research from operational contexts. Even carefully done large sample, predictive validation studies may not provide a good indication of how well the test will work after it is transitioned for use in employment screening. In the early 1980s, an empirically keyed Army biodata instrument measuring several temperament constructs, called the Military Applicant Profile, was carefully developed and validated on a large research sample. When transitioned to the operational setting, its predictive validity quickly declined to near zero, coupled with substantial score inflation (White et al., 2001).

Similarly, in the late 1990s, the Army evaluated AIM in a preimplementation research program, in which they conducted a predictive validation with over 20,000 recruits in addition to simulated applicant faking experiments. When AIM was subsequently implemented as an operational screening measure in 2000, we observed large changes in test scores and validity. Compared to the research sample, those in the operational sample scored significantly higher on AIM and their scores were less predictive of attrition (Young et al., 2004). These changes were not anticipated from results obtained from our research samples, including our simulated applicant faking experiments.

These boundaries are neither impermeable nor rigid. By understanding their nature and their limitations, we can expand the utility of our measures. We have found that some items are valid in a concurrent design but overestimate predictive relationships, whereas others are valid in both concurrent and longitudinal designs. Our results indicate that the average concurrent validities of items high in job content (e.g., questions about how hard you typically work) can be inflated, simply for the reason that they ask individuals to disclose behaviors that pertain directly to their performance on their current job. In concurrent validation studies, responses to personality items high in job content seem to function more like self-ratings of job performance and thereby overestimate the predictive validities (White & Moss, 1995). Accordingly, we have found that items conceptually linked to performance on the job, yet rated as low in job content by expert panels, are more likely to have similar validity coefficients in concurrent and predictive contexts.

The divergence between results obtained in a research setting from those obtained in an operational setting is magnified when the stakes are high and the number of participants is large. However, we have found that it is possible to develop items that are effective even across highly distinctive settings. We have had some success using less transparent, but criterion-valid items, that show greater resistance to deliberate faking. We are also exploring the use of forced-choice response formats, in combination with item-response theory methods, to reduce fakability (Stark, Chernyshenko, & Drasgow, 2005).

https://doi.org/10.1111/j.1754-9434.2008.00049.x Published online by Cambridge University Press

## Closing Comments

In spite of these challenges, we have had some success in incorporating personality measures into the Army personnel selection and classification system. Our results indicate that the use of personality measures for employee screening can have substantial economic utility by reducing costly turnover, improving job performance, or expanding recruiting markets. Importantly, considerable test utility is possible even when the criterion-related validities are relatively low (White et al., 2001; Young & White, 2006). In some applications, we have increased selection utility by combining personality measures with cognitive ability tests to achieve higher criterion-related validity than is possible when either is used separately (White et al., 2004).

Accordingly, we recommend continuing to investigate self-report measures of personality and approaches to reduce, detect, and counter the unintended, negative effects of faking on hiring decisions and criterion-related validity. These include: (a) developing improved methods for detecting socially desirable response tendencies; (b) using warnings with consequences that faking can be detected; (c) applying scoring adjustments to eliminate any hiring bias favoring high fakers, without lowering criterion-related validity; (d) incorporating measures of response distortion and their interactions with personality constructs into equations for predicting job success; (e) retesting when faking is elevated; and (f) using other validated sources of information for evaluating candidates identified as high fakers.

We are encouraged by this progress and see several important avenues for future research. First, we see a need to continue to explore alternatives to our current method of measuring personality, particularly for use in high-stakes testing. Just as Hough and Oswald spoke of the promise of a facet-based approach, we see great potential in an item-response theory–based (Stark et al., 2005) and facet-based approach to personality measurement (Roberts, Chernyshenko, Stark, & Goldberg, 2005). Not only do we believe this approach will result in a more valid test, but it also enhances the opportunity for development of multiple forms, computer-adaptive testing, and improved control of faking.

Second, research is needed to gain a better understanding of the various applicant situations and factors that moderate the generalizablity of validities from job incumbents to applicants. We believe that the mind-set and motivation of test takers greatly varies both within and between research and applicant settings. Before we can make substantial improvements in simulating the actual applicant environment, a better understanding of the variability in applicant mindsets and the situations that affect this is crucial.

Finally, we believe that the field could be served by hearing more from the researchers who have worked to transition personality tests to operational use. We have gone far beyond the stage where the value of personality testing was disparaged and believe we are on the threshold of an era that will recognize such testing as an essential component of the selection process.

## References

Hough, L. M., Eaton, N. L., 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 [Monograph]. *Journal of Applied Psychology, 75*, 581–595.

Roberts, B. W., Chernyshenko, O. S., Stark, S., & Goldberg, L. R. (2005). The structure of conscientiousness: An empirical investigation based on seven major personality questionnaires. *Personnel Psychology, 58*, 103–139.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The multi-dimensional pairwise preference model. *Applied Psychological Measurement, 29*, 184–201.

White, L. A., Hunter, A. E., & Young, M. C. (2006, May). Social desirability effects on the predictive validity of personality constructs. In R. Griffith & Y. Yoshita (Chairs), *Deceptively clear: Applicant faking behavior and the prediction of job performance*. Symposium conducted at the 21st Annual Conference of the Society for Industrial-Organizational Psychology, Dallas, TX.

White, L. A., & Kilcullen, R. N. (1998, April). How socially desirable responding affects the criterion-related validity of self-report measures of personality.

https://doi.org/10.1111/j.1754-9434.2008.00049.x Published online by Cambridge University Press

In M. McDaniel (Chair), *Applicant faking with non-cognitive tests: Problems and solutions*. Symposium presented at the 13th Annual Meeting of the Society for Industrial and Organizational Psychology, Dallas, TX.

White, L. A., & Moss, M. C. (1995, April). Factors influencing the concurrent versus predictive validities of personality constructs. In F. Schmidt (Chair), *Response distortion and social desirability in personality testing for personnel selection*. Symposium presented at the 10th Annual Conference of the Society of Industrial and Organizational Psychology, Orlando, FL.

White, L. A., Young, M. C., Heggestad, E. D., Stark, S., Drasgow, F., & Piskator, G. (2004). *Development of a non-high school diploma graduate pre-enlistment screening model to enhance the Future Force*. Paper presented at the 24th Annual Conference of the Army Science Conference, Orlando, FL.

White, L. A., Young, M. C., & Rumsey, M. G. (2001). ABLE implementation issues and related research. In J. P. Campbell & D. J. Knapp (Eds.), *Exploring the limits in personnel selection and classification* (pp. 525–558). Hillsdale, NJ: Erlbaum.

Young, M. C., & White, L. A. (2006, November). *Preliminary operational findings from the Army's Tier Two Attrition Screen (TTAS) measure*. Paper presented at the 25th Army Science Conference, Orlando, FL.

Young, M. C., White, L. A., Heggestad, E. D., & Barnes, J. D. (2004, July). *Operational validation of the Army's new pre-enlistment screening measure*. Paper presented at the Annual Conference of the American Psychological Association, Honolulu, HI.

https://doi.org/10.1111/j.1754-9434.2008.00049.x Published online by Cambridge University Press