Detecting Faking on a Personality Instrument Using Appropriateness Measurement

Michael J. Zickar and Fritz Drasgow

University of Illinois, Urbana-Champaign

Research has demonstrated that people can and often do consciously manipulate scores on personality tests. Test constructors have responded by using social desirability and lying scales in order to identify dishonest respondents. Unfortunately, these approaches have had limited success. This study evaluated the use of appropriateness measurement for identifying dishonest respondents. A dataset was analyzed in which respondents were instructed either to answer honestly or to fake good. The item response theory approach classified a higher number of faking respondents at low rates of misclassification of honest respondents (false positives) than did a social desirability scale. At higher false positive rates, the social desirability approach did slightly better. Implications for operational testing and suggestions for further research are provided. *Index terms: appropriateness measurement, detecting faking, item response theory, lying scales, person fit, personality measurement.*

Much literature has linked measures of personality traits with behavior in organizational settings. For example, Sparks (1983) found consistent relationships between scores on a standardized personality scale and measures of job success, job effectiveness, and management potential. Personality variables such as conscientiousness and anxiety have been found to correlate with absenteeism and turnover (Bernardin, 1977). Army enlistees who were low in traits such as emotional stability and nondelinquency had a higher drop-out rate during a four-year army term (White, Nord, Mael, & Young, 1993). Barrick & Mount (1991) discovered a small but consistent relationship ($r = .22$) between conscientiousness and a wide variety of criteria across a broad range of jobs in a meta-analysis of previous research. Extroversion was also a significant predictor of job-related behaviors for both sales and management positions ($r = .15$ and $r = .18$, respectively). Although these relationships are lower than validity coefficients typical of cognitive ability tests, personality measures assess quite different human attributes, thus providing incremental validity when combined with cognitive ability measures. Given this potential, personality constructs would be expected to be prevalent in personnel selection programs. However, companies have been reluctant to include personality instruments in their programs; instead, they primarily use ability tests and interviews. One reason for this reluctance is the possibility of faking on personality measures.

Past research has established that respondents are able to significantly distort scores on a wide variety of personality measures (e.g., Gillis, Rogers, & Dickes, 1990; Krahe, 1989). Respondents who are instructed to answer personality measures in a pattern that will present themselves in a favorable light typically receive higher scores than respondents instructed to answer honestly or than those given no instructions. Thus, it seems clear that personality scales can be consciously manipulated. However, there is some disagreement on the prevalence of faking in real-life operational situations.

In an Army sample, Hough, Eaton, Dunnette, Kamp, & McCloy (1990) compared respondents who had no motivation to distort responses with actual applicants and found similar scores between groups. Contrary to those findings, Anderson, Warner, & Spector (1984) found that almost half of the job applicants for

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 20, No. 1, March 1996, pp. 71–87
© Copyright 1996 Applied Psychological Measurement Inc.
0146-6216/96/010071-17$2.10

a variety of positions claimed that they had experience performing at least one of several imaginary tasks that the researchers had invented, such as "matrixing solvency files." Job applicants who claimed experience on these spurious tasks also inflated their responses on items related to experience on real tasks. Regardless of the prevalence of faking, organizations will continue to resist using personality measures in operational programs in which errors have a high cost, as long as the potential to fake with impunity exists.

Detecting Faking in Personality Measurement

One approach to countering faking is to write items that are difficult to fake. Becker & Colquitt (1992) found that respondents distort less on personality items for which the answers can potentially be verified. For example, a question such as "Do you enjoy talking to people?" is difficult to verify, but the question "Were you a member of any social clubs while in high school?" could be verified by school records, yearbooks, and so forth. According to Becker & Colquitt (1992), there would be more faking on the former, less objective item.

Another approach is to ask questions that are ambiguous or less transparent about what is being measured (Edwards, 1970). These subtle items usually are generated by an external validation procedure; items are selected that have mean differences between different groups. Unfortunately, there are problems with the external validation technique: (1) items may function differently in cross-validation samples, and (2) research has indicated that subtle items often have lower validity than more transparent items (e.g., Burkhart, Gynther, & Fromuth, 1980; Duff, 1965; Wiener, 1948).

If personality instruments cannot easily be made resistant to faking, then an alternative solution would be to identify those respondents who have distorted their responses. Research on this approach goes back to the 1940s with the seminal work on the Minnesota Multiphasic Personality Inventory (MMPI). The MMPI was originally designed for psychodiagnostic evaluation (Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989). Three scales designed to detect invalid responses were embedded within the MMPI (Meehl & Hathaway, 1946). Scale F was composed of 64 items that were answered with an extremely low frequency in one direction by a normal sample. For example, a respondent who answers false to "I believe in law enforcement" (the key was formed in the 1940s) as well as answering other items in a similarly unlikely manner would score high on the F Scale. The MMPI K Scale uses items that differentiated patients with known psychological disorders whose MMPI profiles appeared normal and respondents with no psychological disorders. Gough (1950) proposed an additional detection index composed of an individual's score on the F scale minus the score on the K scale (i.e., F – K). One additional scale, the MMPI L Scale consists of items that have a socially desirable answer that cannot honestly be answered in the extreme direction by more than a small number of individuals. An example is "I read all the editorials in the newspaper every day." If a number of these questions are answered in the affirmative, there is reason to believe that the respondent is not answering honestly. The L Scale was rationally constructed, as opposed to the construction of the other two scales, which used purely empirical methods. Many detection scales used on personality inventories, often called social desirability scales, are rationally constructed, such as the L Scale.

Numerous studies (e.g., Bagby, Buis, & Nicholson, 1995; Gillis et al., 1990; Gough, 1950; Lanyon, 1993) have examined the usefulness of detection scales for identifying individuals who consciously distort responses. Much of this research has been conducted using the MMPI. Of the work done on the MMPI, a primary concern has been in detecting respondents who are exaggerating mental symptoms in order to attract attention (i.e., faking bad). For instance, Gillis et al. (1990) used cut scores on the F – K index that had in previous research differentiated between normal individuals and psychiatric patients responding honestly versus normal individuals feigning psychopathological symptoms. The recommended cut score of the F – K index identified 92% of the fakers but misclassified 13% of the honest sample. Similarly, Gough (1950) found that 58% of the faking normals were correctly identified, and 1% of the honest sample was misclassified using the

F – K index. Thus, there appears to be some utility in using the F – K score in identifying individuals explicitly manipulating responses. However, it is difficult to generalize from detection of individuals feigning mental illnesses (or its lack) to the detection of normal individuals faking good on personality items.

Lanning (1989) suggested that it may be possible to detect normal respondents who are faking good. Lanning computed a regression equation using scores on a "good impression" scale along with other substantive personality scales on the California Personality Inventory to differentiate between a sample of college students asked to fake good and a heterogeneous sample of normal individuals. The scores derived from the regression equation achieved a hit rate of 67% with a 1% false positive (FP) rate. In a faking context, hit rate refers to the percentage of fakers correctly classified with a particular cut score on a detection index; the FP rate refers to the percentage of honest respondents incorrectly identified as fakers with that same score. However, the generalizability of the results based on a regression equation is difficult to determine when the two samples also differ in composition. The high hit rate at such a low FP rate may suggest that the regression equation also differentiated between college students and others, a feature that aided differentiation of fakers in this context but that would be irrelevant in other situations. Thus, the effectiveness of social desirability scales in detecting faking in an organizational context is still somewhat unclear.

Detecting Other Kinds of Unusual Responses

There has also been research developing scales designed to detect response patterns that appear to be random. The variable response inconsistency scale (VRIN; Butcher et al., 1989) was designed for detecting inconsistent responses on the MMPI. Wetter, Baer, Berry, Smith, & Larsen (1988) administered the MMPI under four experimental conditions. Respondents were instructed to either answer honestly, simulate a moderate psychological disturbance, simulate a severe psychological disturbance, or answer randomly. In the random response condition, respondents filled out the answer sheet without access to the questionnaire items. Although VRIN identified individuals in the random response condition, there were not mean differences between individuals in the simulated psychological disturbance conditions and the honest condition. Thus, this scale may be useful for detecting individuals who are responding to items in a manner that suggests lack of comprehension, misgridding (e.g., an individual who misgrids an optical scanning sheet by answering Item 10 in the Item 11 blank and continues answering in the wrong blanks), or idiosyncratic personality trait structures (e.g., see Reise & Waller, 1993; Waller & Reise, 1992). VRIN seems, however, to have little power to detect intentional faking.

Another strategy used to detect distorted responses is based on response latencies. It has been hypothesized that respondents who distort their responses take a longer time to respond to individual items presented on a computer (Hsu, Santelli, & Hsu, 1989). This strategy may have limited practical value because respondents with low dexterity or unfamiliarity with computers may have a higher rate of being classified as fakers. Little research has been conducted using this detection technique.

External procedures for detecting falsified responses use information for the classification decision (i.e., faker vs. honest) that is distinct from the information that is used for the substantive classification (e.g., high vs. low self-esteem). Research on the success of external techniques has been mixed, prompting a search for better techniques. In addition, with external techniques there is the possibility that respondents could be given sophisticated training or coaching to thwart such aberrance classification. Meehl & Hathaway (1946) stated "One may conclude that the intent to deceive is not often detectable by [MMPI Scale] L when the subjects are relatively normal and sophisticated" (p. 538).

Appropriateness Measurement

An internal aberrance detection technique simply uses the information contained in the substantive scale item responses to detect respondents who distort their responses. An internal technique that has developed from item response theory (IRT) may be useful in addressing the problem of faking on personality tests.

Appropriateness measurement (one of a number of procedures for determining person fit) is a technique introduced by Levine & Rubin (1979) to identify mismeasured individuals on a test or scale that provides adequate measurement for a large majority of individuals. For instance, an individual who misgrids an optical scanning sheet will present a confusing pattern of responses with little obvious psychological meaning. Another example would be an examinee who copies a small number of answers from a high ability neighbor when a test administrator leaves the examination room. In the personality testing domain, a respondent who answers verifiable items in an honest manner but answers transparent items in a socially desirable manner will present a seemingly inconsistent pattern of responses that may be possible to identify using appropriateness measurement.

Appropriateness measurement quantifies the difference between an examinee's observed pattern of item responses to responses expected on the basis of that person's standing on the latent trait θ and a set of item response functions (IRFs), as specified by some IRT model. IRFs are functions that relate θ to the probability of affirming an item. An examinee whose pattern of responses greatly differs from the expected pattern of responses will have an extreme appropriateness index.

Levine & Drasgow (1988) developed an approach to optimal statistical analysis for appropriateness measurement. Optimal indexes provide most powerful statistics for detecting aberrant responses. Based on the Neyman-Pearson Lemma, a most powerful statistic uses a likelihood ratio test consisting of the likelihood of a response pattern under a model for aberrant responding and the likelihood of a response pattern given a model for nonaberrant responding. Thus $LR(\mathbf{u}) = P_a(\mathbf{u})/P_n(\mathbf{u})$, where $P_a(\mathbf{u})$ is the likelihood of an observed response pattern $\mathbf{u}$ given a certain model of aberrant responding, and $P_n(\mathbf{u})$ refers to the corresponding likelihood for the nonaberrant model. Models for nonaberrant and aberrant responding should be determined by the characteristics of the test and the nature of the individuals who complete the test or scale.

Purpose

The objective of this study was to examine the effectiveness of IRT appropriateness measurement techniques in detecting respondents who were faking on a personality inventory. Previous work using appropriateness measurement has generally been limited to simulation data because of large sample size requirements and the inherent difficulty of gaining access to an identifiable set of aberrant response patterns. In a noted exception, Reise & Waller (1993) used a practical (i.e., nonoptimal) appropriateness statistic, l_z (Levine & Drasgow, 1982), to identify individuals with seemingly idiosyncratic response patterns on a personality questionnaire. Reise and Waller, however, were not able to judge aberrance classification accuracy because their dataset did not have independently identifiable aberrant response patterns.

In this study, an Army dataset that provided clearly delineated nonaberrant and aberrant samples, each with an adequate sample size, was analyzed. Consequently, the effectiveness of the appropriateness indexes in correctly classifying honest and faking good respondents could be directly tested. Moreover, the IRT approach was compared to a traditional approach to detecting faking good because a social desirability scale was included in the inventory.

Method

ABLE Dataset

The United States Army constructed a large personality inventory as part of its Project A (Peterson, Hough, Dunnette, Rosse, Houston, Toquam, & Wing, 1990). The Assessment of Biographical and Life Events (ABLE) consists of 11 content scales that measure separate personality or temperamental constructs. The ABLE, developed with a factor-analytic approach, was designed to predict attrition in the first term enlistment of new Army recruits. The ABLE includes a social desirability scale (SoD; 13 items), designed to detect respondents who answer questions in a socially desirable fashion (i.e., faking good). Respondents who chose the most socially desirable option received a score of 1 for that item; all other options were given a 0 score. Respondents with high scores on this scale might be asked to verify answers.

Six substantive scales were selected from the ABLE for analysis in the present study: Emotional Stability (ES; 17 items), Cooperativeness (COOP; 18 items), Nondelinquency (NOND; 20 items), Work Orientation (WO; 19 items), Internal Control (IC; 16 items), and Energy Level (EL; 21 items).

Datasets

Two datasets from a large-scale Army research project were made available for this research (White et al., 1993). The first dataset, hereafter called the *validation* dataset, consisted of 48,725 respondents who were administered the ABLE inventory by paper-and-pencil upon entrance to the U.S. Army. Respondents were told that personnel decisions (e.g., promotion or dismissal) would not be based on ABLE scores.

The second set of respondents ($N = 1{,}987$) took part in an experimental study with several conditions: Respondents were instructed either to answer in a fashion that would make them "look good" or to answer honestly. All examinees in this experiment were informed that their responses would not be used in future personnel decisions. $N = 324$ respondents were asked to answer all questions honestly (the *honest* condition). Two *fake good* conditions were investigated. In one condition, respondents ($N = 550$) were asked simply to present themselves in a "good light" (the *adlib faking* condition). In the second condition, respondents ($N = 550$) were asked to present themselves in a "good light" and then were coached on how to respond to items in a fashion that would present themselves in a "good light" (the *coached faking* condition). The coaching consisted of feedback on three practice items.

Models

Recent research has examined the use of IRT models, developed in the context of ability testing, for personality assessment (e.g., see Drasgow & Hulin, 1990; Muraki, 1990; Reise & Waller, 1990, 1993; Waller & Reise, 1992). The two-parameter logistic model (2PLM) has been used extensively on personality and attitude scales because of its simplicity and attractive properties for this type of data. This model has been demonstrated to provide reasonable fit to personality data (Reise & Waller, 1990).

The 2PLM is a model for dichotomously scored responses and it incorporates an item response function (IRF), which denotes the probability of selecting the positively keyed option given θ. The 2PLM has the form

$$P(u_i = 1|\theta = t) = \frac{1}{1+\exp[-1.7a_i(t-b_i)]}, \tag{1}$$

where

a_i is the discrimination parameter for item i, $i = 1, ..., n$,

b_i is the location parameter for item i,

u_i is the response of the person with trait level θ to item i, and

1.7 is a scaling constant.

Because the 2PLM is for dichotomous responses, if there are more than two answer options in an item, there must be an artificial dichotomization so that one or more options is recoded to be the single positive category and all remaining options are recoded to a single negative category. Because the ABLE has three options, the first two options were negatively keyed (i.e., scored 0) and the third option was positively keyed (i.e., scored 1). Although the 2PLM has the advantage of simplicity, some important information may be lost when polytomous responses are dichotomized. Therefore, the data were also analyzed with polytomous IRT models.

Polytomous models may be a more appropriate class of models for accurately representing personality data. This class of models retains all information from the item options because no dichotomization is required. Polytomous models may provide additional information beyond dichotomous models when the negatively keyed options and the positively keyed option reflect varying levels of the underlying trait.

Two different models for polytomous responses were used: Samejima's (1969) graded response model (GRM) and Bock's (1972) nominal model (NM). Although IRFs characterize dichotomous IRT models, polytomous models use option response functions (ORFs) to relate θ to the probability of endorsing a particular option. For the GRM and NM, location parameters are associated with particular options of items. The GRM assumes an a priori ordering among the options, which seems appropriate for the ordered option format of the ABLE. The probability of selecting option h on item i is

$$P(v_i = h|\theta = t) = \frac{1}{1+\exp\left[-1.7a_i\left(t-b_{i,h}\right)\right]} - \frac{1}{1+\exp\left[-1.7a_i\left(t-b_{i,h+1}\right)\right]}, \tag{2}$$

where v_i denotes person i's response to the polytomously scored items and h is the particular option selected by the response ($h = 1, ..., s_i$, where s_i refers to the number of options for item i).

For the GRM, all options within an item are assumed to have equal as. b is allowed to vary within the constraint $b_{h-1} < b_h < b_{h+1}$. Muraki (1990) used this model successfully to model actual responses to Likert items from a personality scale.

The NM is more flexible than the GRM because there is no a priori ordering of options. Here, the probability of selecting option h on item i is written for the s_i options of this item. Note that the scaling constant 1.7 has been absorbed into other terms to simplify notation:

$$P(v_i = h|\theta = t) = \frac{\exp\left(a_{i,h}t + b_{i,h}\right)}{\sum_{h'=1}^{s_i}\exp\left(a_{i,h'}t + b_{i,h'}\right)}. \tag{3}$$

The NM allows the a and b parameters to vary for each option within an item. The additional flexibility in this model should be advantageous when large sample sizes are available.

Analyses

Honest responding condition. The fit of the models for normal (i.e., honest) responding was investigated. The six ABLE scales were selected for analysis based on scale length—longer scales were selected because they should provide more information about the traits assessed. Research has also shown that longer scales provide better detection rates of simulated aberrant responses (Reise & Due, 1991).

A spaced sample of 3,000 examinees was selected from the dataset of 48,725 respondents by sampling every 16th response pattern (beginning with the first response vector) until 3,000 patterns had been selected. This subsample was used to fit the nonaberrant respondent model. Consistent with previous analyses of the experimental datasets (Young, White, & Oppler, 1991) and validation datasets (White et al., 1993), there was an average of one-fifth to one-fourth of a standard deviation (SD) difference in means for data collected under validation conditions and data collected in the honest experimental condition. Because these differences were relatively small, the models estimated from the validation sample seemed to provide a reasonable approximation for responding under the honest experimental condition.

The NM and GRM were estimated with marginal maximum likelihood estimation using the MULTILOG program, version 6.0 (Thissen, 1991). The 2PLM was estimated with marginal maximum likelihood estimation using the PC-BILOG program, version 3 (Mislevy & Bock, 1991). 30 quadrature points were used in estimation. Omitted responses were not included in the likelihood equation.

The fit of each IRT model was evaluated in an independent cross-validation sample of 3,000 respondents, again drawn from the complete dataset by the procedure used to form the calibration sample, but starting with the second response vector. χ^2 statistics were computed in the cross-validation sample by dividing the n scale items into $n/3$ sets of three items. Scale items were grouped into bundles of three items that varied in endorsement rate. χ^2 statistics were computed for all single items, the three item pairs formed by all the combinations of items within a bundle, and the item triple. Item bundles were formed to reduce the number of item doubles and triples and, hence, reduce computation time. Computing χ^2 statistics for item doubles and triples was necessary for assessing the fit of potential interactions between items within a scale. Expected proportions of respondents selecting each option were computed by

$$P(v_i = h) = \int P(v_i = h|\theta) f(t)\mathrm{d}t, \tag{4}$$

where f is the density of θ, which was taken as the standard normal. The logic is similar for dichotomous items. The expected proportions for pairs of items, say items i and j, were computed by

$$P(v_i = h,\, v_j = k) = \int P(v_i = h|\theta) P(v_j = k|\theta) f(t)\mathrm{d}(t). \tag{5}$$

The expected proportions for triples of items follow the same logic. In sum, χ^2 statistics for single items, pairs of items, and item triples were computed using expected proportions computed from the calibration sample item parameter estimates and observed proportions from the cross-validation sample. The χ^2 statistics were divided by their degrees of freedom (df) to provide an index of model fit.

Fake good conditions. Next, a model for the respondents in the *fake good* conditions was conceptualized. The aberrant model was operationalized as a single θ shift in which faking occurred only on items that were transparent (Drasgow, Williams, Mead, Levine, Thomasson, & Tsien, 1990). Therefore, for each item that was deemed fakeable respondents were hypothesized as responding as a person with a θ level shifted +.50 to the right on the θ scale. On other items, the aberrant model assumed individuals would answer similarly to those in the honest condition (i.e., no θ shift). Thus, using the 2PLM as an example, fakeable items were modeled as

$$P(u_i = 1|\theta = t) = \frac{1}{1 + \exp\{-1.7a_i[(t + .50) - b_i]\}}. \tag{6}$$

Equations for the other models were similarly modified. Several levels of θ shift (e.g., +.25 and +1.0) were analyzed but provided virtually identical results.

Items were deemed fakeable if there was a significant mean difference on that item between a sample of 32 honest examinees and 55 coached fake good examinees drawn from the experimental sample. These respondents were sampled by taking every 10th response pattern from both conditions. This resulted in $N = 491$ in the coached faking sample, $N = 291$ in the honest sample, and $N = 550$ in the adlib faking sample. The examinees used for this analysis were not included in the actual classification study to avoid any capitalization on chance. Because item descriptions were not available (for security reasons), it can only be hypothesized that less fakeable items were more objective and less socially desirable than the more fakeable items.

Detection of aberrant responding. This study compared two types of optimal appropriateness indexes—the single scale index (LR_x) for scale x and the multiscale index ($LR_{x,y}$) for scales x and y—to the SoD scale in their power to detect faking respondents. Both types of optimal indexes used the likelihood ratio developed by Levine & Drasgow (1988) with the response pattern likelihood for the 2PLM, GRM, and NM as the denominator and the response pattern likelihood for the θ-shift modification of the denominator as the numerator. $LR_{x,y}$, developed by Drasgow, Levine, & McLaughlin (1991), combines the information from two separate unidimensional scales to provide additional power in detecting aberrant responses. (The software used was limited to the analysis of two scales.) This technique is particularly useful if individual scales are short, which was the case with the ABLE.

Index effectiveness was examined with receiver operating characteristic (ROC) functions in the two faking conditions. In the coached faking condition, the three indexes (LR_x, $LR_{x,y}$, and the SoD scale) were used to distinguish respondents in the experimental honest condition from respondents in the experimental coached faking condition. In the adlib faking condition, these indexes were used to distinguish respondents in the experimental honest condition from respondents in the adlib faking condition. To compute ROC functions for an index, response vectors from the honest condition and the relevant faking condition were sorted in ascending order on scores of the index. At each specific value on LR_x, $LR_{x,y}$, and the SoD scale, respondents with scores greater than that cut score were classified as faking; respondents with lower scores were classified as honest. The proportion of fakers correctly classified (hits) and the proportion of honest respondents misclassified as aberrant (FPs) were calculated at each cut score. Each point on a ROC consists of a hit rate (plotted as the ordinate) and a FP rate (plotted as the abscissa).

Rarely will an organization want to use a cut score associated with a FP rate higher than 5%. Therefore, the percentage of fakers correctly identified (i.e., hit rate) when up to 5% of the honest respondents were incorrectly classified as faking (i.e., FPs) was examined. For a given IRT model, there were 42 ROC functions computed [six single scale indexes (LR_x) and 15 multiscale indexes ($LR_{x,y}$) of all possible pairs of single scales were computed for each of the two faking conditions]. In order to simplify the presentation of these ROC functions, hit rates were computed at six FP rates (.5%, 1%, 2%, 3%, 4%, and 5%). Because points on the ROC functions were rarely associated with FP rates exactly at the six desired values, linear interpolation based on the two closest observed data points was used to calculate the hit rate at the particular FP rate. For some of the FP rates, there were not any observed data points close to the FP, so that the linear interpolation probably was not accurate. For example, in order to compute the hit rate at the FP rate of 1% for $LR_{\text{NOND,IC}}$ for the coached condition, linear interpolation of the observed data points (FP rate, hit rate) of (0.00000, 0.00000) and (.01712, .34343) was used to estimate the value (.01000, .20060). Linear interpolation underestimates hit rates because ROC functions are generally not linear at low FP rates (see Drasgow et al., 1991). Hit rates that were estimated without at least one data point within .005 of the estimated FP rate were labeled poor estimates.

Results

Scale Characteristics

Table 1 shows the coefficient α internal consistency estimates for the six ABLE scales and the SoD scale, computed in the sample of 3,000 respondents used for model estimation. Also, the number of items for which there was a significant mean difference between the subsamples of coached fake good and honest examinees is given for the six scales that were used in the analyses.

Table 1
Number of Items, Number of Mean Differences, and Cronbach's α for the ABLE Scales

Scale	Number of Items	Number of Mean Differences	Cronbach's α
Emotional Stability (ES)	17	9	.84
Cooperativeness (COOP)	18	7	.81
Nondelinquency (NOND)	20	3	.79
Work Orientation (WO)	19	9	.84
Internal Control (IC)	16	5	.80
Energy Level (EL)	21	13	.85
Social Desirability (SoD)	13		.63

The six scale scores (from the sample of 3,000 respondents used for model estimation) were moderately correlated with each other (Table 2). The correlations ranged from .37 to .75, with a median of .52.

t tests for each item on the six scales were computed to test the differences between the coached faking and honest conditions. An item with a *t* value significant at $p < .05$ was classified as significant. Because the power of this test was limited by the relatively small samples of the two groups drawn from the experimental samples ($N = 55$ for the coached faking sample; $N = 32$ for the honest sample), an experiment-wise error rate correction was not used. This procedure for identifying fakeable items classified from 15% to 62% (with a mean of 41%) of the items as fakeable per scale (see Table 1).

Table 2
Intercorrelations of ABLE Scales

	Scale					
Scale	ES	COOP	NOND	WO	IC	EL
ES	—					
COOP	.55	—				
NOND	.37	.56	—			
WO	.52	.54	.47	—		
IC	.46	.47	.43	.51	—	
EL	.69	.55	.43	.75	.55	—

Respondents in the coached faking condition generally had substantive scale scores that were 1 SD higher compared to respondents in the honest condition. Respondents in the adlib faking condition had scores that averaged one-half SD higher than the honest condition. Therefore respondents in the coached faking condition should be more easily identified.

The unidimensionality of the scales (required by the 2PLM, NM, and GRM) was supported by factor analysis. Examinations of scree plots indicated that all scales had one dominant first factor underlying the data. However, the COOP scale had a second factor that had an associated eigenvalue of 1.21.

IRT Results

Summary statistics for the *a* and *b* parameter estimates of the 2PLM are shown in Table 3. Note that the estimates of *a* were not particularly high.

Table 4 shows that the 2PLM generally fit the data better than either of the polytomous models. The 2PLM had lower item single χ^2 means than the NM for all six scales and lower means than the GRM for five of the six scales. Some of these differences were substantial (e.g., for ES: 1.25 for 2PLM, 3.25 for NM, and 4.05 for GRM), while others were minor (e.g., for COOP: 3.80 for 2PLM, 3.85 for NM, and 3.52 for GRM). For item doubles, the 2PLM had lower χ^2 means than both the NM and GRM for all six scales. For item triples, the 2PLM had lower χ^2 means than the NM for only two scales and lower means than the GRM for three scales. Differences between

Table 3
Mean and SD of 2PLM Item Parameters for ABLE Scales

	a		*b*	
Scale	Mean	SD	Mean	SD
ES	.69	.23	.30	.60
COOP	.64	.16	–.07	.72
NOND	.59	.19	.28	1.48
WO	.81	.24	.29	.64
IC	.84	.30	–.66	.69
EL	.71	.16	.22	.82

Table 4
Frequency, Mean, and SD of χ^2 to df Ratios for ABLE Scales

Scale, Model,	Frequency Distribution of χ^2 to df Ratio					
and Items	<1	1–2	2–3	>3	Mean	SD
Emotional Stability						
2PLM						
Singles	12	1	0	4	1.25	1.56
Doubles	6	6	2	2	2.09	2.51
Triples	1	2	1	1	2.63	2.03
NM						
Singles	3	5	2	7	3.25	2.60
Doubles	0	1	3	12	4.33	1.72
Triples	0	0	0	5	3.80	.85
GRM						
Singles	3	4	2	8	4.05	3.75
Doubles	0	1	1	14	6.30	3.05
Triples	0	0	0	5	5.14	1.99
Cooperativeness						
2PLM						
Singles	5	4	2	7	3.80	4.02
Doubles	1	2	2	13	5.52	4.51
Triples	0	0	1	5	5.58	2.70
NM						
Singles	3	4	2	9	3.85	2.96
Doubles	0	2	5	11	5.77	3.61
Triples	0	0	1	5	5.20	1.28
GRM						
Singles	2	6	2	8	3.52	2.64
Doubles	0	1	4	13	5.91	3.54
Triples	0	0	1	5	5.36	1.64
Nondelinquency						
2PLM						
Singles	7	7	2	4	2.36	2.91
Doubles	4	3	2	10	3.10	1.99
Triples	1	0	1	4	3.31	1.91
NM						
Singles	7	3	2	8	2.84	2.32
Doubles	1	4	3	11	3.30	1.45
Triples	0	1	4	1	2.77	.72
GRM						
Singles	5	5	1	9	2.87	2.25
Doubles	1	4	2	12	3.36	1.74
Triples	0	1	3	2	2.85	.82
Work Orientation						
2PLM						
Singles	13	1	2	3	1.85	3.59
Doubles	2	2	4	10	4.99	6.27
Triples	0	1	1	4	6.45	6.42
NM						
Singles	6	4	4	6	2.37	3.53
Doubles	0	3	5	10	5.47	7.62
Triples	0	0	1	5	5.62	4.65

continued on the next page

Table 4, continued
Frequency, Mean, and SD of χ^2 to *df* Ratios for ABLE Scales

Scale, Model, and Items	Frequency Distribution of χ^2 to *df* Ratio <1	1–2	2–3	>3	Mean	SD
Work Orientation, continued						
GRM						
Singles	6	5	1	7	2.85	2.25
Doubles	0	3	3	12	5.87	7.00
Triples	0	0	0	6	5.83	4.43
Internal Control						
2PLM						
Singles	8	3	2	3	1.69	1.96
Doubles	6	5	0	4	2.53	2.73
Triples	1	1	0	3	2.79	1.51
NM						
Singles	4	9	1	2	1.80	1.93
Doubles	3	3	6	2	2.61	1.70
Triples	0	2	1	2	2.35	.81
GRM						
Singles	4	6	3	3	2.89	3.67
Doubles	0	6	3	5	5.30	5.76
Triples	0	2	1	2	4.34	3.45
Energy Level						
2PLM						
Singles	14	2	2	5	2.02	3.32
Doubles	4	4	6	7	3.31	3.40
Triples	0	1	3	3	3.73	2.89
NM						
Singles	6	8	2	5	2.27	2.42
Doubles	0	3	5	13	4.52	3.01
Triples	0	1	1	5	4.17	1.62
GRM						
Singles	6	6	4	5	2.55	2.56
Doubles	1	2	4	14	4.87	3.21
Triples	0	1	1	5	4.30	1.57

the two polytomous models were minimal, although the NM fit the data better on five of the six scales. The χ^2 to *df* ratio statistics obtained under the 2PLM were comparable to the values obtained by Drasgow, Levine, Tsien, Williams, & Mead (1995) who fit several polytomous models to five standardized achievement tests. However, the χ^2 values obtained under the NM and GRM were slightly higher than those obtained by Drasgow et al. (1995). The χ^2 values for pairs and triples of items generally exceeded the values for single items across all models. This is in contrast to Drasgow et al. (1995) who found that χ^2s for single items were generally larger than χ^2s for item pairs and item triples. Also, the item single χ^2s for all models were elevated for the COOP scale and for the ES scale modeled with the polytomous models.

Identification of Aberrant Responding

Tables 5 and 6 present the observed hit rates and FP rates for the SOD scale and the appropriateness measurement index based on the 2PLM (tables showing the results for the two polytomous models are available from the first author). Table 5 provides results for when the coached condition was designated as the aberrant sample; Table 6 provides similar results for the adlib faking good condition.

Across faking conditions there were large differences in detection rates for both the IRT and social desirability techniques. At a FP rate of 5% the average hit rate for $LR_{x,y}$ for the 2PLM was 53% in the coached condition and 26% in the adlib condition. The hit rate for the SoD scale at the 5% FP rate was 62% in the coached condition and 28% in the adlib condition. This difference between conditions slightly increased when the FP rate was decreased so that at the .5% FP rate the average hit rate for $LR_{x,y}$ was 30% in the coached condition compared to only 9% in the adlib condition. The average hit rate for the social desirability approach at the .5% FP rate was 14% in the coached condition and only 1% in the adlib condition.

There were differences for the effectiveness of appropriateness measurement across the six substantive scales. The WO and EL scales had the highest hit rates across all FP rates and both faking conditions. The ES scale had the next highest hit rates for almost all FP rates across conditions and models. The COOP and NOND scales had the next highest hit rates, except in the adlib faking condition where the NOND scale did not have sufficient data across FP conditions. Similarly, the IC scale did not have sufficient data in either condition, probably because there were only five fakeable items on that scale.

The multiscale index, $LR_{x,y}$, provided a moderate increase in detection rates compared to the indexes based on single scales, especially at low FP rates. In the coached sample, the increase in $LR_{x,y}$ hit rates (averaged across all scales or possible extensions) over the single scale index ranged from 2% to 133% larger than single scale hit rates. In the adlib sample, the increase in $LR_{x,y}$ hit rates ranged from 0% to 40%.

Table 5
Percent of Hits for FP Rates From 5% to .5% in the Coached Condition Using the 2PLM

Appropriateness	FP Rate					
Index and Scale	5%	4%	3%	2%	1%	.5%
Single Scales (LR_x)						
ES	51%	48%	46%	38%*	24%	12%*
COOP	45%	41%	40%	32%	16%*	8%*
NOND	45%	44%	43%	32%*	16%*	8%*
WO	53%	47%	45%	44%	40%	20%*
IC	20%*	16%*	12%*	8%*	4%*	2%*
EL	57%	55%	50%	39%	33%	23%
Average (All)	45%	42%	39%	32%	22%	12%
Average (Good Est.)	50%	47%	45%	38%	32%	23%
Multiscale ($LR_{x,y}$)						
ES, COOP	55%	52%	47%	43%	40%	30%
ES, NOND	58%	55%	48%	41%	35%	30%
ES, WO	61%	53%	50%	47%	38%	27%
ES, IC	54%	51%	49%	40%	39%	32%
ES, EL	60%	59%	56%	50%	32%	24%
COOP, NOND	55%	51%	46%	39%	31%	30%
COOP, WO	57%	55%	51%	45%	43%	33%
COOP, IC	47%	42%	39%	36%	32%	16%*
COOP, EL	56%	55%	54%	51%	43%	31%
NOND, WO	59%	57%	52%	49%	33%	32%
NOND, IC	52%	51%	49%	37%	20%*	10%*
NOND, EL	60%	58%	52%	48%	44%	20%
WO, IC	58%	54%	51%	47%	36%	34%
WO, EL	58%	56%	54%	50%	47%	39%
IC, EL	57%	53%	48%	46%	39%	30%
Average (All)	53%	50%	46%	45%	37%	28%
Average (Good Est.)	53%	50%	46%	45%	38%	30%
Social Desirability	62%	55%	53%	47%	29%	14%

*Entry was poorly estimated due to lack of data in that range.

Table 6
Percent of Hits for FP Rates From 5% to .5% in the Adlib Condition Using the 2PLM

Appropriateness	FP Rate					
Index and Scale	5%	4%	3%	2%	1%	.5%
Single Scales (LR_x)						
ES	23%	19%	18%	15%	13%	8%
COOP	21%	20%	18%	15%	12%	7%
NOND	5%*	4%*	4%	3%*	1%*	1%*
WO	26%	22%	20%	19%	17%	8%
IC	12%*	9%*	7%*	5%*	2%*	1%*
EL	27%	26%	22%	16%	13%	10%
Average (All)	19%	17%	15%	12%	10%	6%
Average (Good Est.)	24%	22%	16%	16%	14%	8%
Multiscale ($LR_{x,y}$)						
ES, COOP	26%	24%	18%	16%	15%	10%
ES, NOND	22%	20%	15%	3%	3%	2%*
ES, WO	30%	22%	21%	20%	14%	9%
ES, IC	23%	21%	20%	17%	13%	9%
ES, EL	28%	27%	25%	21%	11%	8%
COOP, NOND	23%	15%	12%	2%*	2%*	2%*
COOP, WO	29%	26%	24%	19%	16%	11%
COOP, IC	21%	18%	16%	14%	11%	6%*
COOP, EL	22%	19%	18%	15%	11%	6%
NOND, WO	26%	25%	17%	15%	2%*	2%
NOND, IC	16%*	8%*	4%*	3%*	1%*	1%*
NOND, EL	29%	25%	18%	12%	11%	2%
WO, IC	32%	29%	27%	20%	13%	12%
WO, EL	28%	26%	24%	22%	16%	14%
IC, EL	28%	24%	20%	18%	13%	11%
Average (All)	26%	22%	19%	14%	10%	7%
Average (Good Est.)	26%	23%	20%	16%	12%	9%
Social Desirability	28%	22%	16%	12%	5%	1%

*Entry was poorly estimated due to lack of data in that range.

The IRT approaches had higher hit rates than the social desirability approach at low FP rates. However, at higher FP rates, the social desirability approach had slightly higher hit rates. Using the 2PLM and multiscale ROC function (with good estimates), the average increase in detection rates over the SoD scale in the coached sample was 114% at a .5% FP rate, 31% at the 1% rate, –4% at the 2% rate, –13% at the 3% rate, –9% at the 4% rate, and –15% at the 5% rate. In the adlib condition, the increases in hit rates at the same set of FP rates were 800%, 140%, 33%, 25%, 5%, and –7%.

There were no large differences in detection rates across any of the IRT models. Rarely was there a difference in detection rate greater than 5% between any of the models for a given FP rate. Thus, the additional information supplied by the polytomous models apparently did not lead to higher hit rates.

Discussion

This study is unique in the appropriateness measurement literature. Previous research has generally used simulation data, real data with aberrance artificially created by the researcher, or real data without any identifiable aberrant sample. This study used real data with some examinees distorting their own responses. Therefore, this study demonstrated the limits and advantages of appropriateness measurement in real-data conditions.

There is some difficulty in translating results from the personality domain to applications in ability testing. With the research by Reise and Waller (Reise & Waller, 1990, 1993; Waller & Reise, 1992) and Muraki (1990) providing a notable exception, it is nonetheless true that much less research has been devoted to developing IRT models for personality data. For example, unidimensional IRT models seem well suited for fitting ability test data in which there is usually one correct answer with other answers possibly varying in their degree of correctness. In personality measurement, the trait that is assessed may not be as simply conceptualized as verbal or mathematical skill in ability testing. Further, many personality scales have not been subjected to the rigorous factor-analytic development that was used for the development of the scales used in this study. Thus, some personality scales may be more subject to problems of multidimensionality than are ability tests. Although there was a dominant first factor for each of the scales, the COOP scale had slight multidimensionality, which is likely the explanation for the poor fit of the unidimensional IRT models for this scale. In addition, the COOP scale had slightly lower hit rates in the coached faking good analysis than other scales with similar length and a similar number of items deemed fakeable. Both the ES and WO scales had approximately 50% fakeable items and both scales had higher detection rates throughout the range of FP rates. However, in the adlib faking analysis the hit rates for the COOP scale were not substantively lower than hit rates for the other scales. Drasgow, Levine, Williams, McCusker, Thomasson, & Lim (1989) found substantial reductions in detection of simulated spurious high and low aberrant responses when a unidimensional model was used to analyze multidimensional items. More simulation research needs to be conducted to examine the effects of differing levels of multidimensionality on detection of aberrant responses.

Another reason that results may differ between ability and personality testing is the nature of aberrance itself. In ability testing, the most obvious way to fake good is to copy an answer from a high ability neighbor or to memorize answers based on advance information. In personality scales, the "correct" answer is very often transparent, so dissimulation seems more likely than in ability testing.

Presumably there should be a certain ratio of faking-resistant items to fakeable items in order to provide the best detection in personality measurement. With too few fakeable items, fakers will be difficult to detect. The two scales in the present study with the lowest percentage of fakeable items (NOND and IC) consistently had low hit rates. Conversely, with too many fakeable items, it will be difficult to differentiate the high honest respondents from the fakers.

Some of the ROC analyses computed from appropriateness indexes for scales with few fakeable items (e.g., the NOND scale) often did not have data points at FP rates below the 3% rate. For example, the NOND scale had only three items out of 20 that were deemed fakeable. The likelihood values using the aberrant model will not be very different from the likelihood values under the normal model because only three items will differ. Therefore, there will be many honest individuals who will be classified as aberrant (at any hit rate) because their responses to those few items (in this case, three) did not fit the model. It seems advisable to forego appropriateness measurement on single scales with few transparent items. However, $LR_{x,y}$ appeared to solve this problem.

A final potential reason for differences between results on personality and cognitive ability instruments could be the prevalence of *untraited* individuals—people who are not adequately represented by a personality trait structure that adequately represents a majority of individuals (Reise & Waller, 1993; Tellegen, 1988). If there were a significant number of untraited individuals in the ABLE dataset, results might have been distorted. Future research should examine the prevalence of untraited individuals in real datasets.

Overall, the detection power of both the SoD scale and the appropriateness measurement indexes was too low to justify use in most operational situations. One explanation for the low power of the appropriateness measurement indexes was that the discrimination parameters were lower than typically used in simulation studies. Meijer, Molenaar, & Sijtsma (1994) demonstrated that lower *a* parameters in the 2PLM resulted in lower detection rates of simulated aberrant responses.

Examinees in the adlib faking condition were much more difficult to detect than those in the coached condition. This is consistent with a finding by Drasgow et al. (1991) that fakers are easier to detect when faking is more prevalent and, in this case, extreme.

Polytomous models did not provide better aberrance detection than dichotomous models. A potential explanation is that there were many items that had a low percentage of respondents selecting the extremely low option. The ORFs for those items may have been poorly estimated compared to the other options. Thus, it appears there may have been little information available to the polytomous models for these personality scales. Clearly, polytomous models should only be used when the data are adequate.

This research suggests that if test users are committed to detecting fakers, there are several reasons they should consider using appropriateness measurement techniques instead of, or in addition to, traditional social desirability scales. First, if FPs are extremely costly or embarrassing to a test administrator, then the IRT approach may provide substantial benefits by minimizing unnecessary retesting and possible bad feelings generated by misclassifying honest individuals. This conclusion seems warranted because the increases in aberrance detection over the traditional social desirability approach were substantial in the critical low FP rates. Second, an internal method of detection may be more appealing than external techniques. For example, potential examinees could be trained to be aware of items obviously high in social desirability. This coaching may occur commonly for instruments providing critical information in a selection procedure. Third, internal methods do not require extra items in the measuring instrument.

There are some obvious limitations to this study. First, there was only one social desirability scale administered to each individual, so comparisons between IRT approaches and social desirability approaches would benefit from more research using different social desirability scales. Second, the model of aberrance in this study was formulated based on its simplicity. Undoubtedly the nature of aberrance is more complex than a simple uniform θ shift on fakeable items. For example, a reviewer commented that an individual with a relatively low θ should fake more than an individual with relatively high θ. Future research should evaluate the utility of more complex models of aberrance. Third, the method for determining which items were fakeable was not ideal. A better approach would have been to classify items based on a content analysis. However, capitalization on chance was minimized by not including the examinees used in the classification analysis. Finally, the software was limited to using only two scales for $LR_{x,y}$. With additional work, optimal multitest appropriateness indexes can be generalized to larger numbers of unidimensional scales.

References

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

Bagby, R. M., Buis, T., & Nicholson, R. A. (1995). Relative effectiveness of the standard validity scales in detecting fake-bad and fake-good responding: Replication and extension. *Psychological Assessment, 7,* 84–92.

Barrick, M. R., & Mount, M. K. (1991). The big-five personality dimensions job performance: A meta-analysis. *Personnel Psychology, 44,* 1–26.

Becker, T. E., & Colquitt, A. L. (1992). Potential versus actual faking of a biodata form: An analysis along several dimensions of item type. *Personnel Psychology, 45,* 389–406.

Bernardin, H. J. (1977). The relationship of personality variables to organizational withdrawal. *Personnel Psychology, 30,* 17–27.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29–51.

Burkhart, B. R., Gynther, M. D., & Fromuth, M. E. (1980). The relative predictive validity of subtle versus obvious items on the MMPI depression scale. *Journal of Clinical Psychology, 36,* 748–751.

Butcher, J., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *Minnesota Multiphasic Personality Inventory 2: Manual for administration and scoring.* Minneapolis: University of Minnesota Press.

Drasgow, F., & Hulin, C. L. (1990). Item response theory. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (Vol. I; pp. 577–636). Palo Alto CA: Consulting Psychologists Press.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1991). Appropriateness measurement for some multidimensional test batteries. *Applied Psychological Measurement, 15,* 171–191.

Drasgow, F., Levine, M. V., Williams, B., McCusker, C., Thomasson, G. L., & Lim, R. G. (1989). *An evaluation of optimal appropriateness measurement for use in practical settings* (AFHRL-TP-89-41). Brooks Air Force Base TX: Air Force Human Resources Laboratory.

Drasgow, F., Levine, M. V., Tsien, S., Williams, B., & Mead, A. D. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement, 19,* 143–165.

Drasgow, F., Williams, B., Mead, A., Levine, M. V., Thomasson, G. L., & Tsien, S. (1990). *Identifying unmotivated examinees. Final Report: Contract DAAL03-86-D-0001.* San Diego CA: Navy Personnel Research and Development Center.

Duff, F. L. (1965). Item subtlety in personality inventory scales. *Journal of Consulting Psychology, 29,* 565–570.

Edwards, A. L. (1970). *The measurement of personality traits by scales and inventories.* New York: Holt, Rinehart and Winston.

Gillis, J., Rogers, R., & Dickes, S. (1990). The detection of faking bad response styles on the MMPI. *Canadian Journal of Behavioural Science, 22,* 408–416.

Gough, H. G. (1950). The F minus K dissimulation index for the Minnesota Multiphasic Personality Inventory. *Journal of Consulting Psychology, 14,* 408–413.

Hathaway, S. R., & McKinley, J. C. (1983). *The Minnesota Multiphasic Personality Inventory manual.* New York: Psychological Corporation.

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.

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

Krahe, B. (1989). Faking personality profiles on a standard personality inventory. *Personality and Individual Differences, 10,* 437–443.

Lanning, K. (1989). Detection of invalid response patterns on the California Psychological Inventory. *Applied Psychological Measurement, 13,* 45–56.

Lanyon, R. I. (1993). Development of scales to assess specific deception strategies in the Psychological Screening Inventory. *Psychological Assessment, 5,* 324–329.

Levine, M. V., & Drasgow, F. (1982). Appropriateness measurement: Review, critique, and validating studies. *British Journal of Mathematical and Statistical Psychology, 35,* 42–56.

Levine, M. V., & Drasgow, F. (1988). Optimal appropriateness measurement. *Psychometrika, 53,* 161–176.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4,* 269–290.

Meehl, P. E., & Hathaway, S. R. (1946). The K factor as a suppressor variable in the MMPI. *Journal of Applied Psychology, 30,* 525–561.

Meijer, R. R., Molenaar, I. W., & Sijtsma, K. (1994). Influence of test and person characteristics on nonparametric appropriateness measurement. *Applied Psychological Measurement, 18,* 143–151.

Mislevy, R. J., & Bock, R. D. (1991). *BILOG users' guide.* Chicago: Scientific Software.

Muraki, E. (1990). Fitting a polytomous item response model to Likert type data. *Applied Psychological Measurement, 14,* 59–71.

Peterson, N. G., Hough, L. M., Dunnette, M. D., Rosse, R. L., Houston, J. S., Toquam, J. L., & Wing, H. (1990). Project A: Specification of the predictor domain and development of new selection/classification tests. *Personnel Psychology, 43,* 247–276.

Reise, S. P., & Due, A. M. (1991). The influence of test characteristics on the detection of aberrant response patterns. *Applied Psychological Measurement, 15,* 217–226.

Reise, S. P., & Waller, N. G. (1990). Fitting the two-parameter model to personality data. *Applied Psychological Measurement, 14,* 45–58.

Reise, S. P., & Waller, N. G. (1993). Traitedness and the assessment of response pattern scalability. *Journal of Personality and Social Psychology, 65,* 143–151.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement,* No. 17.

Sparks, C. P. (1983). Paper and pencil measures of potential. In G. P. Dreher & P. R. Sackett (Eds.), *Perspectives on employee staffing and selection* (pp. 349–368). Homewood IL: Dow-Jones Irwin.

Tellegen, A. (1988). The analysis of consistency in personality assessment. *Journal of Personality, 56,* 621–663.

Thissen, D. (1991). *MULTILOG user's guide.* Chicago: Scientific Software.

Waller, N. G., & Reise, S. (1992). Genetic and environmental influences on item response pattern scalability. *Behavior Genetics, 22,* 135–152.

Wetter, M. W., Baer, R. A., Berry, D. T., Smith, G. T., & Larsen, L. H. (1992). Sensitivity of MMPI-2 validity scales to random responding and malingering. *Psy-*

chological Assessment, 4, 369–374.

White, L. A., Nord, R. D., Mael, F. A., & Young, M. C. (1993). The assessment of background and life experiences (ABLE). In T. Trent & J. H. Laurence (Eds), *Adaptability screening for the armed forces* (pp. 101–162). Washington DC: Office of the Assistant Secretary of Defense.

Wiener, D. N. (1948). Subtle and obvious keys for the Minnesota Multiphasic Personality Inventory. *Journal of Applied Psychology, 12,* 164–170.

Young, M. C., White, L. A., & Oppler, S. H. (1991). *Coaching effects on the Assessment of Background and Life Experiences (ABLE).* Paper presented at the meeting of the Military Testing Association, San Antonio TX.

Acknowledgments

Thanks to Len White and the U.S. Army Research Institute for providing access to the data analyzed in this paper and to Michael Levine and Bruce Williams for their help. Partial support for this research was provided by Office of Naval Research Contract Numbers N00014-89-K-0059 and N00014-90-J-1958; principal investigator, Michael Levine.

Author's Address

Send requests for reprints or further information to Michael Zickar, 603 East Daniels, Champaign IL 61820, U.S.A. Email: mzickar@s.psych.uiuc.edu.