

Detection of Aberrant Responding on a Personality Scale in a Military Sample: An Application of Evaluating Person Fit With Two-Level Logistic Regression

Carol M. Woods and Thomas F. Oltmanns
Washington University in St. Louis

Eric Turkheimer
University of Virginia

Person-fit assessment is used to identify persons who respond aberrantly to a test or questionnaire. In this study, S. P. Reise's (2000) method for evaluating person fit using 2-level logistic regression was applied to 13 personality scales of the Schedule for Nonadaptive and Adaptive Personality (SNAP; L. Clark, 1996) that had been administered to military recruits ($N = 2,026$). Results revealed significant person-fit heterogeneity and indicated that for 5 SNAP scales (Disinhibition, Entitlement, Exhibitionism, Negative Temperament, and Workaholism), the scale was more discriminating for some people than for others. Possible causes of aberrant responding were explored with several covariates. On all 5 scales, severe pathology emerged as a key influence on responses, and there was evidence of differential test functioning with respect to gender, ethnicity, or both. Other potential sources of aberrancy were carelessness, haphazard responding, or uncooperativeness. Social desirability was not as influential as expected.

Keywords: person fit, aberrant responding, personality assessment, personality disorders

Supplemental materials: <http://dx.doi.org/10.1037/1040-3590.20.2.159.supp>

Person fit is the degree to which an item response model fits for an individual examinee (Meijer, 1996; Meijer & Sijtsma, 1995, 2001; Reise, 1995; Reise & Flannery, 1996; Tellegen, 1988). For example, in a general sample of people, the probability of responding "true" to items about personality or psychopathology is expected to decrease as items become more severe (i.e., as these items measure more extreme personality or psychopathology). Thus, in the general population, fewer people should endorse an item such as "I avoid important activities because of my anxiety" (more extreme anxiety) than an item such as "I get nervous before speaking in front of an audience" (less extreme anxiety). This pattern of decreasing endorsement with increasing severity, also called consistency (Tellegen, 1988) or scalability (Reise & Waller, 1993), should hold for each person on all scale items.

For dichotomously scored items, a perfectly scalable individual endorses all items with severity below his or her level of the latent variable, θ , and endorses no items with higher severity. This endorsement pattern is also called a perfect Guttman scale (Guttman, 1950). For example, when items are ordered by severity (e.g., $-2.52, -1.50, -1.32, -1.17, -1.08, -1.05, -0.98, -0.18, 0.05, 0.18, 0.20, 0.47, 0.79, 1.18, 1.20, 1.70, 1.80, 3.91$), the perfectly scalable pattern of responses for someone with $\theta = 1.10$ is 11111111111100000 (1 = endorsement). The more responses deviate from this pattern of decreasing endorsement with increas-

ing severity, the poorer the person fit. Poor person fit can be caused by, for example, carelessness, malingering, or uncooperativeness. See Reise and Waller (1993) for examples of more and less scalable response patterns observed empirically for the Social Closeness scale of the Multidimensional Personality Questionnaire (Tellegen, 1982).

Many methods for the assessment of person fit have been developed (see review by Meijer & Sijtsma, 2001). In the present study, we illustrate one method introduced by Reise (2000) that appears to have been applied to empirical data only once before (Woods, 2008). The method is somewhat complicated, which may be why it has been infrequently applied. As with most person-fit methods, one first estimates item parameters and each person's θ using item response theory (IRT). Reise's (2000) method requires subsequent application of two-level logistic regression. A key advantage is that the approach permits not only identification of aberrant responding but modeling of hypothesized causes of person-fit heterogeneity. Continuous or categorical person-level covariates are easily added to the model to explain the heterogeneity.

In the research presented here, we used Reise's (2000) method to test for aberrant responding to true/false self-report items about personality pathology on 15 trait and temperament scales that form the core of the Schedule for Nonadaptive and Adaptive Personality (SNAP; Clark, 1996). If significant person-fit heterogeneity was observed on a SNAP scale, we have included several person-level covariates to help identify sources of the aberrancy.

Two-Level Logistic Regression for Person Fit

The Person Response Function

In Reise's (2000) approach, logistic regression is used for estimation of the parameters of a person response function (PRF;

Carol M. Woods and Thomas F. Oltmanns, Department of Psychology, Washington University in St. Louis; Eric Turkheimer, Department of Psychology, University of Virginia.

Correspondence concerning this article should be addressed to Carol M. Woods, Psychology Department, Campus Box 1125, Washington University in St. Louis, St. Louis, MO 63130. E-mail: cwoods@artsci.wustl.edu

Lumsden, 1978; Nering & Meijer, 1998; Sijtsma & Meijer, 2001; Trabin & Weiss, 1983; Weiss, 1973, as cited in Sijtsma & Meijer, 2001). The PRF shows the relationship between item endorsement and item severity for each person. In contrast to an item response function (IRF), which indicates how well an item discriminates among persons with higher and lower levels of θ , a PRF indicates how well a person discriminates among items of varying severity. An IRF shows how the probability of responding “true” depends on θ for one item, whereas a PRF shows how the probability of responding “true” depends on item severity for one person.

Figure 1 plots the PRF for one of the best fitting and one of the worst fitting response patterns we observed for the SNAP Exhibitionism scale. The line with dots as the plotting symbol has a large negative slope (-2.22 , intercept = -0.11), which indicates decreasing endorsement with increasing severity. With items ordered by estimated severity parameters ($\beta_i = -1.58, -1.05, -0.79, -0.79, -0.56, -0.29, -0.16, -0.15, -0.12, 0.03, 0.16, 0.3, 0.43, 0.53, 0.72, 1.34$), this man’s responses were 1111111100000000; his θ was estimated to be -0.06 . The other line in Figure 1 (with crosses as the plotting symbol) shows the PRF for a more aberrant responder, for whom the relation between endorsement and severity was weak (PRF slope = -0.37 , intercept = -0.58). For this woman, the response pattern was 0101101000000101 and the estimated θ was -0.31 .

As is explained subsequently, β_i and θ are estimated (in the same metric) with IRT, and PRF slopes and intercepts are estimated with logistic regression. We plotted these example PRFs using the logistic regression equation solved for the probability that the response (u_{ij}) to item i from person j is “1,”

$$p(u_{ij} = 1) = \frac{\exp(b_0 + b_1 x_i)}{1 + \exp(b_0 + b_1 x_i)}, \quad (1)$$

with the slope (b_1) and intercept (b_0) parameters given above and β_i used as the predictor (x_i).

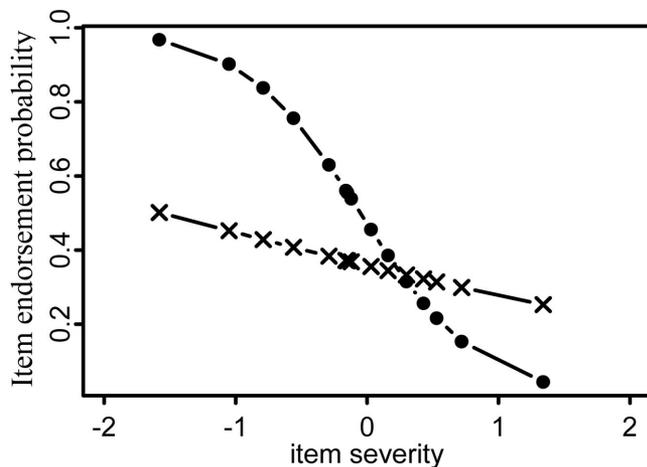


Figure 1. Two empirical person response functions (PRFs) for the SNAP Exhibitionism scale. The curve with dots for the plotting symbol shows a strong inverse relationship between item endorsement probability and item severity (PRF slope = -2.22). The curve with crosses for the plotting symbol shows a weak endorsement-severity relationship (PRF slope = -0.37). Each dot or cross corresponds to an estimated severity parameter for each of the 16 Exhibitionism items.

To understand the range of magnitude for a PRF slope, we can exponentiate the range and interpret in terms of odds ratios. For the best fit pattern above, the odds of item endorsement are multiplied by $\exp(-2.22) = 0.11$ for every one-unit increase in item severity. For the worst fit pattern, the odds of item endorsement are multiplied by $\exp(-0.37) = 0.69$ for every one-unit increase in item severity. One unit of item severity is one standard deviation of θ . When there is absolutely no relation between endorsement and severity, the PRF slope is 0, and the odds of item endorsement are unchanged as severity increases [$\exp(0) = 1$].

The odds of responding “1” versus “0” for any particular value of β_i can be obtained by plugging that value into

$$\exp(b_0 + b_1 \beta_i)$$

Because $\exp(0) = 1$, the odds of responding “1” versus “0” are equal when

$$\beta_i = \frac{-b_0}{b_1}$$

For patterns that match the model, the odds are equal when $\beta_i = \theta$. For the best fit pattern above, the odds of “1” versus “0” are equal when $\beta_i = -0.05$, which is nearly equal to this respondent’s estimate of $\theta(-0.06)$. These odds do not hold for patterns with poor fit: For the worst fit pattern above, the odds are equal when $\beta_i = -1.57$ (estimated θ was -0.31).

If the PRF slope estimate is about the same for all people, then either the scale discriminates fairly well and about equally for all individuals or the scale has poor measurement properties, so that aberrant responding cannot be detected. Scales discriminate better when the items are highly related to one another and to θ and when values of β_i are spread over a relatively wide range of θ s. If a scale lacks these good properties for most people, aberrant responding is not detectable.

If a scale has good properties for most people, it is possible to identify persons for whom the scale is less discriminating. If PRF slopes vary significantly over persons, as they did for the Exhibitionism scale, the scale is not equally discriminating for all individuals; thus, the item response model does not fit equally well for all persons. Individuals for whom fit is worst are identified on the basis of their estimated PRF slope (slopes closest to 0 are worst). Reise’s (2000) approach stands out among PRF-based methods, because it provides sound methods for the estimation of PRF slopes for individuals.

IRT Precedes Logistic Regression

Using Reise’s (2000) method, one fits an item response model to the data prior to the logistic regression analyses to estimate the severity (β_i) of each item (i.e., threshold parameter) and θ for each person. This method constitutes a typical application of IRT (see introductions by Embretson & Reise, 2000, or Thissen & Wainer, 2001) and can be carried out with standard software, such as BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2003) or MULTILOG (Thissen, 1991).

In the present analysis, we use a variation of classic IRT to obtain estimates of β_i and θ . Implicit in classic IRT is the assumption that θ is normally distributed in the population of people. This is unlikely to be true for all variables. Ramsay-curve IRT (RC-

IRT; Woods, 2006a, 2007; Woods & Thissen, 2006) can be used for estimation of the θ distribution simultaneously with the item parameters. If the latent distribution is approximately normal, the answers match those from standard software, such as MULTILOG. If the distribution is not approximately normal, RC-IRT provides more accurate item parameter estimates than do methods that assume normality (Woods, 2006a, 2007; Woods & Thissen, 2006). RC-IRT is used for the present analyses.

Two-Level Logistic Regression

In general, logistic regression is a variant of linear regression that is adapted to be appropriate for a categorical (binary, ordinal, or nominal) outcome variable. In the study reported here, we employed logistic regression for a binary outcome (refer to Agresti, 1996, 2002; Collett, 2003; or Hosmer & Lemeshow, 2000). In the models to be fitted, each true/false response to a SNAP item is an outcome, and the β_i of that item is used as the predictor. The coefficient from this regression is the PRF slope that reflects the relationship between item response and item severity.

However, the PRF slope is potentially different for different people. A logistic regression model that permits the PRF slope to vary over individuals (i.e., that permits it to be treated as random rather than fixed) has two levels (references on multilevel models include Raudenbush & Bryk, 2002, and Snijders & Bosker, 1999). The first level is the regression of item response on item severity described above, with the slope treated as random. The error variance of this slope indicates the degree to which it varies over persons.

If there is statistically significant variability in PRF slopes, the PRF slope can be treated as an outcome in a second level of the model, with person-level predictors entered to explain its variability. However, only the systematic portion of its variability is explainable. A reliability coefficient that is a function of the Level 2 prediction error variance and the Level 1 residual variance (Woods, 2008, gives details) will be used to approximate the proportion of variability in the PRF slopes that could potentially be explained by covariates. Reliability coefficients are interpreted as the proportion of the total variance that is reliable.

The intercept from the Level 1 regression may also vary over persons. In the present context, the random intercept is analogous to θ , so it is expected to vary significantly over people and to be completely explainable by IRT scores. If it behaves otherwise, this is an indication that some of the IRT assumptions are violated (see Reise, 2000, or Woods, 2008, for further explanation). No evidence of this type of violation appeared in our analyses, so the intercept is not discussed further.

Covariates

Gender and ethnicity will be entered as covariates, because systematic differences in responding are commonly observed among demographic groups. If group membership is a strong predictor of person fit, this can be interpreted as evidence of differential test functioning (DTF; Shealy & Stout, 1993). DTF occurs when a scale has different measurement properties for one group versus another, controlling for true mean differences in the trait being measured. When differences in measurement properties occur for individual items, it is called differential item functioning

(DIF; Camilli & Shepard, 1994; Holland & Wainer, 1993; Millsap & Everson, 1993).

Sometimes, even if many items function differently between groups, DIF cancels out at the scale level, so DTF is trivial or nonexistent. Thus, DIF does not always imply DTF. If one is interested in scale-level functioning, it makes sense to test for DIF only when there is evidence of DTF. In the present research, DTF is tested as part of the more general person-fit assessment. Evidence of DTF implies DIF, so, if we find DTF, subsequent analyses beyond the scope of the present research will be warranted to identify differentially functioning items.

Another set of covariates includes scores on three SNAP validity scales, which were designed to identify certain sources of invalid responding. Although it was unclear what types of invalidity to expect, there were probably special demand characteristics associated with our sample of Air Force recruits who were tested at the end of basic military training. They were told that their SNAP responses would not be shared with the Air Force. Nevertheless, they may have worried that their scores would influence the next step of their careers. These circumstances may have been especially likely to elicit socially desirable responding.

Additionally, it was unclear how seriously recruits responded to the task, so there may have been aberrancy due to carelessness, uncooperativeness, or haphazard responding, and some recruits may have suffered from a distressed mood or personality pathology. Prior to basic training, recruits were screened for obvious mental disorders by the military, but there was no professionally administered standardized battery that screened for major mental disorders. After training (when recruits filled out the SNAP), neither military screening nor professionally administered standardized assessment was carried out. Recruits who were seriously disturbed could have had elevated scores on the SNAP validity scale designed to detect malingering.

If higher validity-scale scores are associated with more aberrant PRF slopes, the source (e.g., social desirability) is implicated as a contributor to poor person fit. However, scores on validity scales do not definitively clarify the source of aberrant responding. Additional information about individuals is needed (Piedmont, McCrae, Riemann, & Angleitner, 2000). For example, researchers must thoroughly evaluate mental health status to distinguish between malingering and true pathology. Haphazard or careless responding to validity-scale items could also produce misleading scores on validity scales. Nevertheless, a significant relationship between person-fit heterogeneity and a particular validity scale would provide a hypothesis about aberrant responding that could be investigated in future research.

The last two covariates were nominations from peers regarding the extent to which each recruit exhibited features of obsessive-compulsive personality disorder (OCPD) or borderline personality disorder (BPD). We examined these specific types of personality disorder because they are likely to have interesting (and opposite) effects on SNAP scores. People who exhibit features of OCPD are preoccupied with orderliness and perfection. People who exhibit features of BPD are markedly impulsive, are emotionally erratic, and have unstable images of themselves and other people. If aberrant responding is partially due to carelessness, recruits with more nominations for features of OCPD might be less likely to give aberrant response patterns. Conversely, if aberrant responding is partially produced by the presence of serious personality pathol-

ogy, recruits with more nominations for features of BPD might be more likely to give aberrant responses.

Serious personality pathology can, of course, manifest in other forms. We could have used peer ratings of antisocial, histrionic, or narcissistic traits, but ratings for these disorders were highly correlated with those for BPD. Multicollinearity among covariates was avoided by the use of just one measure of severe pathology. Furthermore, the extreme impulsivity and disturbed self-image associated with BPD seem particularly relevant to aberrant response patterns that are the topic of these analyses.

The use of peer nominations for pathological personality features allowed us to avoid basing our conclusions about personality features on another self-report measure that would share the same unique perspective or possible biases reflected in participant self-descriptions on the SNAP. Studies of people with and without mental disorders point to the conclusion that there is, at best, only a modest correlation between the ways in which people describe their own personalities and the ways in which they are perceived by others (Biesanz, West, & Millevoi, 2007; Clark, 2007; Watson, Hubbard, & Weise, 2000).

Method

Participants and Data Collection

The sample consisted of 2,026 Air Force recruits (1,265 male, 761 female) who were completing basic military training at Lackland Air Force Base in San Antonio, Texas. Most were between 18 and 25 years of age ($Mdn = 19$ years). They self-identified as White (1,305), African American (348), Hispanic (75), Asian (68), or Native American (17), with the remaining 213 classified as "other." No standardized assessment battery was administered as part of this study to screen for major mental disorders (e.g., schizophrenia, substance use disorders, or mood disorders). However, recruits had been screened by the military for obvious mental disorders when they enlisted and, again, when they entered basic training.

As part of a larger study (see Oltmanns & Turkheimer, 2006; Thomas, Turkheimer, & Oltmanns, 2003), data were collected by computer from demographically heterogeneous training groups called "flights." All recruits are assigned to a flight at the beginning of basic training, and the members of each flight do virtually everything together for 6 weeks. All participants completed self-report items of the SNAP (Clark, 1996) at an individual computer terminal and then rated peers in their flight using the Multisource Assessment of Personality Pathology (MAPP; Oltmanns & Turkheimer, 2006).

The SNAP

Traits and temperaments. The SNAP consists of 375 true/false questions about trait dimensions related to normal personality traits and features of personality disorders. In this analysis, person fit was evaluated for 15 scales that measure traits or temperaments. The temperament scales are Disinhibition (35 items), Negative Temperament (28 items), and Positive Temperament (27 items). The trait scales (and number of items) include Aggression (20), Dependency (18), Detachment (18), Eccentric Perceptions (15), Entitlement (16), Exhibitionism (16), Impulsivity (19), Manipula-

tiveness (20), Mistrust (19), Propriety (20), Self-Harm (16), and Workaholism (18).

Validity. Three SNAP validity scales were used: Rare Virtues (12 items), Deviance (20 items), and Variable Response Inconsistency (VRIN; 22 items). Rare Virtues consists of highly socially desirable behaviors (e.g., "I have never made a promise that I didn't keep") that are rarely true, so high scores often reflect a naive "fake good" response set. Deviance is the opposite: Items reflect severe pathology rarely endorsed by normal participants, so high scores may indicate "faking bad," including malingering. Persons who experience severe pathology also have high Deviance scores (Clark, 1996).

VRIN is designed to identify inconsistent responding while controlling for item content. VRIN consists of pairs of content-matched items keyed in the same direction, so it is inconsistent to answer "true" to one item and "false" to the other (e.g., "People say I drive myself hard" and "I've been told that I work too hard"). High VRIN scores can be due to carelessness, haphazard responding, or uncooperativeness (Clark, 1996). Summed scores (i.e., sums of 0/1 item scores) for each validity scale were used as covariates for the person-fit analyses. We used observed summed scores instead of IRT scores, because it was unclear whether a validity scale was measuring a latent trait.

The MAPP

The MAPP consists of 103 personality traits presented one at a time. Most items are based on features of 10 personality disorders listed in the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.; American Psychiatric Association, 1994); 10 MAPP subscales correspond to these disorders. Recruits were asked to identify at least one member of their flight or training group who exhibited each trait and to indicate the extent to which the target person demonstrated that trait (0 = *never*, 1 = *sometimes*, 2 = *usually*, 3 = *always*). Item scores for each target are the mean rating over judges, and scale scores are the mean of item scores. Scores for the BPD and OCPD scales were used for the present research.

Item Response Analysis

Although the two-parameter-logistic (2PL; Birnbaum, 1968) IRF is theoretically appealing for personality items, the three-parameter-logistic IRF (3PL; Birnbaum, 1968) sometimes fits significantly better than does the 2PL (e.g., Reise & Waller, 2003). The 3PL includes a third parameter for each item (g), which is the lower asymptote of the IRF and interpreted as a guessing parameter for items with correct answers. For personality items for which endorsement indicates more of a positive trait, g is sometimes interpretable as a social desirability parameter (Zumbo, Pope, Watson, & Hubley, 1997). Reise and Waller (2003) discussed alternative interpretations for positive items and items for which endorsement indicates more of a negative or pathological trait.

In a simulation study, Woods (in press) found that the likelihood ratio chi-square test that compared the 2PL and 3PL IRFs with the θ distribution fixed at normal correctly pointed toward the 3PL, even when the θ distribution was actually nonnormal. This finding informed the analytic strategy used in the present study. For each SNAP scale, the 2PL and the 3PL were compared with normal θ and

Table 1
Means (SDs) of Summed Scores for SNAP Personality and Validity Scales

| SNAP scale | Present sample | | Normative comparison group | |
|------------------------|---------------------|---------------------|----------------------------|---------------------|
| | Men ($n = 1,265$) | Women ($n = 761$) | Men ($n = 281$) | Women ($n = 523$) |
| Aggression | 3.7 (4.0) | 2.9 (3.3) | 6.4 (4.5) | 4.1 (4.1) |
| Dependency | 4.3 (2.9) | 3.9 (3.1) | 4.8 (3.3) | 6.3 (3.7) |
| Detachment | 4.8 (3.8) | 5.0 (3.9) | 6.3 (4.2) | 4.3 (3.9) |
| Disinhibition | 8.8 (5.9) | 7.1 (4.9) | 15.6 (6.3) | 12.1 (6.1) |
| Eccentric Perceptions | 4.8 (3.5) | 4.8 (3.6) | 5.5 (3.5) | 5.2 (3.5) |
| Entitlement | 8.8 (3.3) | 8.7 (3.2) | 9.9 (3.8) | 9.0 (3.6) |
| Exhibitionism | 8.6 (4.0) | 7.8 (3.9) | 9.1 (4.0) | 9.6 (3.7) |
| Impulsivity | 4.6 (3.6) | 4.4 (3.5) | 7.3 (4.1) | 6.7 (4.0) |
| Manipulativeness | 3.6 (3.3) | 2.6 (2.6) | 8.1 (4.2) | 5.7 (4.0) |
| Mistrust | 7.1 (4.4) | 7.7 (4.6) | 7.2 (4.3) | 6.2 (4.3) |
| Negative Temperament | 9.5 (6.5) | 10.7 (6.7) | 13.1 (7.0) | 14.7 (7.0) |
| Positive Temperament | 20.8 (5.0) | 20.4 (4.9) | 19.4 (5.9) | 20.2 (5.8) |
| Propriety | 14.5 (3.3) | 14.6 (3.2) | 11.7 (4.0) | 12.7 (4.0) |
| Self-Harm | 1.2 (1.9) | 1.1 (1.8) | 2.4 (2.8) | 2.1 (2.6) |
| Workaholism | 9.1 (3.6) | 9.2 (3.8) | 8.0 (4.0) | 7.3 (4.0) |
| Validity: Deviance | 1.9 (1.7) | 1.6 (1.6) | 3.5 (2.6) | 2.7 (2.1) |
| Validity: Rare Virtues | 4.9 (2.6) | 5.2 (2.4) | 3.1 (2.1) | 3.2 (2.1) |
| Validity: VRIN | 5.1 (2.2) | 5.0 (2.1) | 6.1 (2.6) | 5.4 (2.3) |

Note. Data for the normative comparison group are from Table 10 (p. 39) of the *Schedule for Nonadaptive and Adaptive Personality (SNAP): Manual for Administration, Scoring, and Interpretation* by Lee Anna Clark. Copyright 1993 by the Regents of the University of Minnesota. All rights reserved. Reproduced by permission of the University of Minnesota Press. "Schedule for Nonadaptive and Adaptive Personality" and "SNAP" are trademarks owned by the University of Minnesota.

SNAP = Schedule for Nonadaptive and Adaptive Personality; VRIN = Variable Response Inconsistency. Sample sizes in header for the comparison group are for the personality scales only. For validity scales, $n_{\text{male}} = 222$ and $n_{\text{female}} = 76$ (Clark, 1996).

RC-IRT (Woods & Thissen, 2006; Woods, 2006a, 2007) was then carried out with the preferred IRF. The 3PL was preferred for a given scale if (a) the chi-square test was significant and (b) the Bayesian information criterion (BIC; Schwarz, 1978) was smaller for the 3PL than for the 2PL. Because the BIC imposes a penalty for the number of parameters estimated, using it with the chi-square test helps to balance parsimony with good fit.

We carried out RC-IRT using the RCLOG (Version 2; Woods, 2006b) program.¹ Expected a posteriori (EAP) estimates of θ for each person and estimates of β_i for each item were saved for use in the logistic regression models. No other item parameters from the IRT analyses were used in subsequent analyses, because PRFs specify relationships only between β_i and item endorsement. Although we used no other item parameters for the two-level logistic regression, the IRF must be specified as accurately as possible, because 2PL-based estimates of β_i are not, in general, equivalent to 3PL-based estimates of β_i .

Person-Fit Analysis

Two-level logistic regression was carried out with maximum likelihood estimation, with standard errors approximated by first-order derivatives (the MLF estimator) implemented in the Mplus program (Version 4.12; Muthén & Muthén, 2006). We computed reliability coefficients with SAS software using formulas given in Woods (2008). In all models, the intercept was treated as a fixed function of EAPs, because, compared with treating the intercept as random, this method improves estimation and convergence (Woods, 2008).

In the first model (Model 1), no predictors of the random PRF slope were included. If statistically significant variability in PRF

slopes was observed in Model 1, which was at least partially systematic (based on the reliability coefficient), a second model (Model 2) that included 11 person-level covariates (described below) was fitted. We estimated PRF slopes for individuals from Model 2 using empirical Bayes methods (Morris, 1983; Snijders & Bosker, 1999, pp. 58–63) implemented in Mplus.

Covariates were scores for the three validity scales, peer-rated BPD and peer-rated OCPD, gender (1 = male, 2 = female), and race. We used five binary variables to code the nominal six-level race variable with White as the reference group. For each SNAP scale, we compared the global fit of models with and without covariates using Akaike's information criterion (AIC; Akaike, 1973) and the BIC. Both are functions of the optimized log likelihood, with a penalty for the number of parameters (AIC) or the number of parameters and the sample size (BIC). Smaller values are preferred.

Results

Descriptive Statistics on SNAP Scales

Means and standard deviations (SDs) of summed scores (i.e., the sum of all 0/1 coded item scores) on each scale are given in Table 1. Normative values from a college sample given in the SNAP manual (Clark, 1996) are listed for comparison. Notice that the military sample is much larger and is about 62% male, whereas the college sample is about 35% male. In general, means for the

¹ The RCLOG computer program is freely available upon request from Carol M. Woods.

military sample were lower on most scales—markedly so on Disinhibition—but were higher on Propriety and Workaholism.

Recruits had just spent an intense, 6-week period in an authoritarian environment in which cleaning and organizing were highly valued activities. People who did not follow the strict rules were likely to be punished by training instructors and chastised by peers. Those who score low on Disinhibition are “serious people who believe in doing things in proper order and following rules of all kinds” (Clark, 1996). Higher scorers on Propriety are “greatly concerned with proper standards of conduct” (Clark, 1996), and higher scorers on Workaholism are “perfectionists . . . who enjoy work more than play” (Clark, 1996). Recruits may have been (a) more perfectionistic and concerned about proper conduct than others when they enlisted, (b) particularly likely to endorse items about these characteristics immediately after basic training, or (c) both.

Another interesting difference between the samples is that military women scored lower than did military men on the Dependency scale but that college women scored higher than did college men (as did college women in another sample described by Oltmanns & Turkheimer, 2006). Perhaps women who self-select into the military and survive basic training tend to be particularly independent. A final observation is that recruits scored higher on Rare Virtues, so they may have been more influenced by social desirability than were members of the normative group. However, the maximum score on Rare Virtues is 12, so the recruits’ scores were not particularly high.

RC-IRT

The 2PL was preferred to the 3PL for all scales. Although 12 of the 15 chi-square tests were significant ($\alpha = .05$), the BIC was always smaller for the 2PL. The difference between BICs for each scale was between 0.18% and 11.04% of the 3PL BIC value.

RC-IRT was carried out using the 2PL for all scales. The θ distribution was approximately normal (skewness [s] = 0; kurtosis [k] = 3) for the majority of the scales but was nonnormal for Dependency ($s = -0.65$, $k = 5.07$), Entitlement ($s = 2.32$, $k = 13.30$), Exhibitionism ($s = 1.30$, $k = 9.14$), and Workaholism ($s = -0.71$, $k = 6.37$). Complete details of the RC-IRT analyses are available upon request from Carol M. Woods but are not given here because they are not the present focus. The purpose of the RC-IRT analysis was to obtain EAPs and estimates of β_i for use in the logistic regression models.

Scales With Poor Item Properties

The majority of items on the Propriety and Manipulativeness scales discriminated poorly. For Propriety, discrimination parameters (a_i s) ranged from 0 to 1.59, and the average over items was 0.81 ($SD = 0.62$). For Manipulativeness, a_i s ranged from 0.18 to 2.31 ($M = 1.35$, $SD = 0.50$). Additionally, all β_i estimates for Manipulativeness were positive; they ranged from 0.61 to 3.49, plus $\beta_{63} = 9.59$. This large severity parameter for Item 63, “People who try to get out of doing something by pretending to need help are probably lazy, not clever,” resulted from the combination of a high endorsement proportion (.86) with poor discrimination ($a_i = 0.18$).

The probability of responding “true” to a weakly discriminating item is about the same for all respondents, so it is impossible even to define aberrancy. It is also difficult to detect aberrancy on scales on which β_i is concentrated in a narrow range of θ , because the restricted range limits the degree to which β_i could possibly covary with endorsement probability. Because aberrancy is difficult to define for poorly discriminating scales and those with narrow-ranging β_i , person fit was not assessed for Propriety and Manipulativeness.

Two-Level Logistic Regression Model 1

Homogeneous PRF slopes for personality scales. The variance of the random PRF slope (b_1) was .00 or .01 and was statistically nonsignificant for eight personality scales. Because the slope did not vary over individuals, the mean (i.e., Level 2 intercept) is a useful summary statistic. Mean PRF slopes were Self-Harm, -1.62 ; Aggression, -1.39 ; Eccentric Perceptions, -1.37 ; Dependency, -1.35 ; Impulsivity, -1.29 ; Mistrust, -1.25 ; Detachment, -1.14 ; and Positive Temperament, -1.12 . The negative slopes are consistent with the usual definition of good person fit: Item endorsement decreased as item severity increased.

Heterogeneous PRF slopes for personality scales. The variance of b_1 was nonzero and was statistically significant for five personality scales: Negative Temperament, Disinhibition, Workaholism, Exhibitionism, and Entitlement. Results from Model 1 are given in Tables 2 and 3, which display indices of global model fit (AIC and BIC), the variance of b_1 (τ) and its standard error (SE), the corresponding reliability coefficient (Λ), and the Level 2 intercept (γ_0) with SE . Reliability coefficients indicated that between 36% and 80% of the person-fit heterogeneity was systematic and could be explained by covariates.

Two-Level Logistic Regression Model 2

Model 2 was fitted to data for the five personality scales that had significant heterogeneity on the basis of Model 1. As shown in Tables 2 and 3, the AIC was smaller for each scale when covariates were included, which indicated better fit. For Exhibitionism, the BIC increased when covariates were added to the model (indicating worse fit), because the BIC rewards parsimony more than does the AIC. However, for every scale, the addition of covariates reduced τ ; this result indicated that the covariates explained at least some of the heterogeneity. Reliability coefficients showed that 19%–50% of the remaining unexplained variance was systematic.

For each scale, Level 2 regression parameters for each covariate (with SE s) are listed in Tables 2 and 3. An asterisk marks those results that significantly predicted PRF slopes ($\alpha = .05$). Below, we describe covariates that significantly predicted aberrant responding. We also mention some nonsignificant effects in which the mean PRF slope was nontrivially more aberrant for Hispanic ($n = 75$) or Asian ($n = 68$) recruits than for White recruits. Though not statistically reliable with present sample sizes, these observations lead to hypotheses to be explored in future research.

Negative Temperament. For Negative Temperament, responses from recruits rated by their peers as higher in BPD were significantly more aberrant: With a 1 SD increase in BPD score, the PRF slope increased by 0.58. Greater aberrancy was also predicted by higher Deviance scores. These two effects suggest

Table 2
Person-Fit Analysis for Negative Temperament, Disinhibition, and Workaholism

| Parameter | Negative Temperament | | Disinhibition | | Workaholism | |
|---------------------------------|----------------------|--------------|---------------|--------------|-------------|-------------|
| | Model 1 | Model 2 | Model 1 | Model 2 | Model 1 | Model 2 |
| AIC | 64,517.18 | 49,432.59 | 67,165.85 | 62,271.35 | 31,558.84 | 31,324.89 |
| BIC | 64,552.96 | 49,566.78 | 67,202.53 | 62,408.88 | 31,592.86 | 31,452.46 |
| τ (variance) | .68 (.04) | .10 (.02) | .39 (.02) | .12 (.01) | .24 (.02) | 0.18 (.02) |
| Λ (reliability) | .80 | .41 | .73 | .47 | .54 | .47 |
| γ_0 (intercept) | -1.28 (.04) | -1.79 (.08) | -1.11 (.03) | -1.05 (.06) | -1.28 (.02) | -2.01 |
| γ_1 (BPD) | — | 0.58* (.20) | — | 0.33* (.14) | — | -0.17 (.18) |
| γ_2 (OCPD) | — | -0.11 (.19) | — | -0.02 (.16) | — | 0.14 (.18) |
| γ_3 (Rare Virtues) | — | 0.02* (.01) | — | -0.05* (.01) | — | 0.01 (.01) |
| γ_4 (Deviance) | — | 0.05* (.01) | — | 0.09* (.01) | — | 0.13* (.01) |
| γ_5 (VRIN) | — | -0.01 (.01) | — | 0.02* (.01) | — | 0.05* (.01) |
| γ_6 (gender) | — | 0.04 (.03) | — | -0.08* (.03) | — | 0.14* (.03) |
| γ_7 (African American) | — | 0.15* (.04) | — | -0.06 (.03) | — | 0.04 (.04) |
| γ_8 (Asian) | — | 0.12 (.08) | — | 0.02 (.06) | — | -0.01 (.10) |
| γ_9 (Hispanic) | — | -0.08 (.10) | — | 0.07 (.05) | — | -0.03 (.09) |
| γ_{10} (Native American) | — | -0.06 (.15) | — | 0.05 (.15) | — | -0.13 (.17) |
| γ_{11} (other) | — | 0.01 (.06) | — | 0.06 (.04) | — | -0.01 (.06) |
| Min, max b_1 | — | -1.96, -0.35 | — | -1.87, 0.28 | — | -2.09, 0.06 |

Note. Values are M (SE). An asterisk indicates that the result is statistically significant ($\alpha = .05$). A dash indicates that the parameter was not estimated. AIC = Akaike's information criterion; BIC = Bayesian information criterion; VRIN = Variable Response Inconsistency; BPD = peer ratings of borderline personality disorder; OCPD = peer ratings of obsessive-compulsive personality disorder; γ = Level 2 regression parameter; b_1 = slope estimate for person response function.

that aberrancy might have been due in part to the presence of pathology. Higher Deviance scores could be due to malingering rather than to true pathology, but malingering seems unlikely in this sample. Higher Rare Virtues scores significantly predicted

Table 3
Person Fit Analysis for Exhibitionism and Entitlement

| Parameter | Exhibitionism | | Entitlement | |
|---------------------------------|---------------|--------------|-------------|--------------|
| | Model 1 | Model 2 | Model 1 | Model 2 |
| AIC | 31,092.36 | 31,009.09 | 30,748.20 | 30,572.81 |
| BIC | 31,125.90 | 31,134.88 | 30,781.74 | 30,698.60 |
| τ (variance) | .24 (.04) | .17 (.04) | .11 (.03) | .04 (.02) |
| Λ (reliability) | .58 | .50 | .36 | .19 |
| γ_0 (intercept) | -1.48 (.03) | -1.89 (.11) | -1.47 (.02) | -1.78 (.09) |
| γ_1 (BPD) | — | 0.70* (.28) | — | 0.24 (.21) |
| γ_2 (OCPD) | — | -0.72* (.27) | — | -0.22 (.21) |
| γ_3 (Rare Virtue) | — | 0.04* (.01) | — | -0.01 (.01) |
| γ_4 (Deviance) | — | 0.10* (.02) | — | 0.12* (.01) |
| γ_5 (VRIN) | — | 0.05* (.01) | — | 0.04* (.01) |
| γ_6 (gender) | — | 0.15* (.05) | — | -0.01 (.04) |
| γ_7 (African American) | — | 0.16* (.07) | — | -0.12* (.05) |
| γ_8 (Asian) | — | 0.19 (.11) | — | 0.24* (.09) |
| γ_9 (Hispanic) | — | 0.16 (.13) | — | -0.05 (.11) |
| γ_{10} (Native American) | — | -0.03 (.22) | — | 0.09 (.34) |
| γ_{11} (other) | — | 0.08 (.08) | — | 0.07 (.06) |
| Min, max b_1 | — | -2.22, 0.31 | — | -2.03, 0.04 |

Note. Values are M (SE). An asterisk indicates that the result is statistically significant ($\alpha = .05$). A dash indicates that the parameter was not estimated. AIC = Akaike's information criterion; BIC = Bayesian information criterion; VRIN = Variable Response Inconsistency; BPD = peer ratings of borderline personality disorder; OCPD = peer ratings of obsessive-compulsive personality disorder; γ = Level 2 regression parameter; b_1 = slope estimate for person response function.

more aberrant PRF slopes, which suggests that socially desirable responding might have produced some aberrancy. Mean PRF slopes were significantly more aberrant for African American versus White recruits and for Asian versus White recruits (the latter difference was not statistically significant). The ethnicity effects can be interpreted as evidence of DTF.

Disinhibition. For Disinhibition, some results were the same as those for Negative Temperament: Aberrancy was significantly greater for recruits with higher Deviance scores and for recruits rated by their peers as higher in BPD. Thus, the presence of pathology was again implicated as a potential cause of aberrant responding. Interestingly, Rare Virtues predicted aberrancy, but the direction of the effect indicated that more aberrancy was accompanied by less socially desirable responding. Higher VRIN scores were associated with greater aberrancy, a result that suggests carelessness, haphazard responding, or uncooperativeness as possible influences. Perhaps one of these influences explains the unexpected direction of the effect for Rare Virtues. The fact that mean PRF slope was significantly more aberrant for men than for women can be interpreted as DTF.

Workaholism. Higher Deviance and VRIN scores predicted greater aberrancy for Workaholism, and the mean PRF slope was significantly more aberrant for female than for male recruits. It is surprising that Rare Virtues appeared to be unrelated to aberrant responding on Workaholism.

Exhibitionism. Many covariates significantly predicted aberrancy on Exhibitionism. As observed for other scales, greater aberrancy was predicted by higher peer-rated BPD and higher Deviance scores. This result implicates pathology as an explanation. Higher VRIN scores and lower peer-rated OCPD were associated with more aberrant PRF slopes, which is consistent with the possibility that aberrancy was partially due to carelessness (and that people with higher levels of OCPD traits were less careless).

Higher scores on Rare Virtues predicted greater aberrancy and suggested that social desirability may have influenced responses. Finally, there was substantial evidence of DTF: The mean PRF slope was more aberrant for women, African Americans, Asians, and Hispanics compared with men and Whites (effects for Asians and Hispanics were not statistically significant).

Entitlement. The finding that higher scores on Deviance and VRIN were associated with greater aberrancy on Entitlement suggests pathology, carelessness, haphazard responding, or uncooperativeness as possible explanations. There was also evidence of DTF: Significant differences in the mean PRF slope indicated greater aberrancy for Asians versus Whites and for Whites versus African Americans.

Correlations Among Person Fit Across Scales

Table 4 lists Pearson correlations among PRF slope estimates for the five scales with significant heterogeneity. Slopes for all scales correlated significantly with one another, though the strength of the relationship varied from .17 to .63. These results indicate that individuals who responded aberrantly to one scale tended to respond aberrantly to the other four scales as well.

Discussion

We used Reise’s (2000) method to evaluate person fit for 13 personality scales from the SNAP (Clark, 1996). Two scales (Propriety and Manipulativeness) were not assessed, because the measurement properties appeared to be poor for most people. Person-fit heterogeneity was observed for 5 of the 13 scales and indicated that the 5 scales discriminated better for some people than for others. Our finding that some heterogeneity was predictable from covariates leads to hypotheses about what might have caused the aberrant responding. Our results are limited by the possibility of measurement error in the scales, covariates, or both, and the explanations we offer are speculative.

A more thorough understanding of the aberrant responding requires additional information from, for example, clinical interviews. Meijer, Egberink, Emons, and Sijtsma (in press) used ancillary information in a study focused on person fit for a scale about children’s self-perceptions: In addition to utilizing additional information about the children’s personal and emotional well-being, they readministered the scale to aberrant responders to check reliability and interviewed teachers about aberrant respond-

ers. In the present study, we explored sources of person fit using peer ratings of BPD and OCPD, scores on three validity scales, gender, and ethnicity. Information provided by peers is particularly useful in this regard, because self-report and informant-report measures provide rather different perspectives on personality pathology (Furr, Dougherty, Marsh, & Mathias, 2007; Miller, Pilkonis, & Clifton, 2005).

Explanations for Aberrant Responding on the SNAP

Severe personality pathology may partly explain aberrant responding. Higher scores on Deviance significantly predicted more aberrant PRF slopes for all five scales. Elevated Deviance scores could indicate feigned pathology. Recruits who were having a particularly difficult time adjusting to the demands of military life may have been trying to “fake bad” in order to find a way out of the commitment they had made when they enlisted.

It is also possible that some recruits suffered from pathology that was not detected by the military during initial screenings or that had been catalyzed by basic training. BPD characteristics, such as extreme impulsiveness, emotional instability, and identity disturbance, were discernible by peers, and higher levels of peer-rated BPD traits predicted aberrant responding for Negative Temperament, Disinhibition, and Exhibitionism. The fact that peer scores for BPD characteristics were found to be associated with aberrant responding on these SNAP scales provides support for the argument that informants are able to provide valid or meaningful data regarding personality pathology, even when those scores are discrepant from the image provided by self-report measures (Fiedler, Oltmanns, & Turkheimer, 2004; Klein, 2003).

Carelessness, haphazard responding, uncooperativeness, or some combination of these influences accounted for some of the aberrancy. Higher VRIN scores were associated with more aberrant PRF slopes for all scales except Negative Temperament. One unique feature of this scale (compared with the other four) is that all constituent items are near the end of the 375-item SNAP (Item 241 and higher). Could recruits have become more careful, deliberate, or cooperative near the end? Alternatively, haphazard responses on the other scales may have been due to confusion, and items on the Negative Temperament scale may have been less confusing. Future research should explore these possibilities.

Scores on Rare Virtues were less related to aberrant responding than we had expected. This may be because Rare Virtues is not a particularly useful measure of social desirability or because the recruits were not as concerned about presenting themselves favorably as we had hypothesized. We used summed scores rather than IRT scores for validity scales, because it was unclear whether a validity scale really measures a latent variable. However, as a sensitivity analysis, we refitted the five 2-level logistic regression models with EAPs rather than summed scores for the validity scales. Rare Virtues EAPs were about as predictive as were summed scores (full results from the sensitivity analyses are available upon request from Carol M. Woods). It seems that the construct validity of validity scales is rarely, if ever, examined empirically (Piedmont et al., 2000). Insightful psychometric evaluations of validity scales would be helpful.

All five personality scales showed evidence of DTF with respect to gender, ethnicity, or both. It would be useful to explore possible causes for these group differences, because there are surely other

Table 4
Pearson Correlations Among PRF Slopes for SNAP Scales With Significant Heterogeneity

| SNAP scale | Disinhibition | Entitlement | Exhibitionism | Negative Temperament |
|----------------------|---------------|-------------|---------------|----------------------|
| Entitlement | .63 | | | |
| Exhibitionism | .33 | .60 | | |
| Negative Temperament | .17 | .29 | .38 | |
| Workaholism | .37 | .62 | .44 | .28 |

Note. N = 2,026. PRF = person response function; SNAP = Schedule for Nonadaptive and Adaptive Personality. All correlations are statistically significant, with p < .0001.

variables (perhaps continuous variables) for which group membership is a proxy. Also, DTF implies DIF, and it would be useful to know which items function differently. Many methods for DIF testing (reviewed by Millsap & Everson, 1993) exist and could be applied to these SNAP scales in future research.

Limitations

Aberrant responding cannot always be detected. Power is best for longer scales that are composed of items that are highly related to one another and to θ and that have relatively widely varying severity levels. Scales with poor measurement properties may appear free of person-fit heterogeneity when, in fact, aberrancy cannot be defined. In the present study, Propriety and Manipulativeness seemed to have poor properties, such that assessing person fit seemed futile. On the basis of our IRT analyses, the other SNAP scales had acceptable measurement properties, so it is encouraging that person-fit heterogeneity was observed for only 5 out of 13 of them. Because person fit is sample dependent, results may differ for samples of college students, patients, community volunteers, or even other samples of Air Force recruits. Replication studies are warranted.

Another limitation inherent in person-fit assessment is that an item response model must first be fitted to data for all respondents. Because it is not possible to know in advance which response patterns are aberrant, they are all combined for IRT. In a simulation study focused on Reise's (2000) method, Woods (2008) found that the presence of aberrant responders produced bias in the IRT results and PRF slopes in a direction that blurred the distinction between aberrant and nonaberrant responders. It would be interesting to test an iterative-purification variation of Reise's method, wherein IRT and two-level logistic regression are repeated several times, with 5% (or some other percentage) of the most aberrant respondents being eliminated each time.

Conclusion

Despite the limitations inherent in person-fit assessment, we believe that Reise's (2000) method can be a useful tool for improving psychological measurement. PRF estimates can help identify people who may have provided invalid data or who may need special attention (e.g., if they suffer from severe pathology). It is sometimes necessary to eliminate aberrant responders from a data set, but this should never be done on the basis of person-fit statistics alone. As much ancillary information as possible is needed if we are to understand what appears to be aberrant responding. Our assessment of the SNAP has raised many questions about variables that may influence response to self-report questionnaires. Exploration of these questions is important, if psychological measurement is to continue to improve.

References

- Agresti, A. (1996). *An introduction to categorical data analysis*. Hoboken, NJ: Wiley.
- Agresti, A. (2002). *Categorical data analysis* (2nd ed.). Hoboken, NJ: Wiley.
- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest, Hungary: Akademiai Kiado.
- American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.
- Biesanz, J. C., West, S. G., & Millevoi, A. (2007). What do you learn about someone over time? The relationship between length of acquaintance and consensus and self–other agreement in judgments of personality. *Journal of Personality and Social Psychology*, *92*, 119–135.
- Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Reading, MA: Addison & Wesley.
- Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.
- Clark, L. (1996). *SNAP manual for administration, scoring, and interpretation*. Minneapolis: University of Minnesota Press.
- Clark, L. (2007). Assessment and diagnosis of personality disorder: Perennial issues and an emerging reconceptualization. *Annual Review of Psychology*, *58*, 227–257.
- Collett, D. (2003). *Modeling binary data* (2nd ed.). Boca Raton, FL: Chapman & Hall.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.
- Fiedler, E. R., Oltmanns, T. F., & Turkheimer, E. (2004). Traits associated with personality disorders and adjustment to military life: Predictive validity of self and peer reports. *Military Medicine*, *169*, 207–211.
- Furr, R. M., Dougherty, D. M., Marsh, D. M., & Mathias, D. W. (2007). Personality judgment and personality pathology: Self–other agreement in adolescents with conduct disorder. *Journal of Personality*, *75*, 629–662.
- Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction* (pp. 60–90). Princeton, NJ: Princeton University Press.
- Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.
- Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: Wiley.
- Klein, D. N. (2003). Patients' versus informants' reports of personality disorders in predicting 7 1/2-year outcome in outpatients with depressive disorders. *Psychological Assessment*, *15*, 216–222.
- Lumsden, J. (1978). Tests are perfectly reliable. *British Journal of Mathematical and Statistical Psychology*, *31*, 19–26.
- Meijer, R. R. (1996). Person-fit research: An introduction. *Applied Measurement in Education*, *9*, 3–8.
- Meijer, R. R., Egberink, I. J. L., Emons, W. H. M., & Sijtsma, K. (in press). Detection and validation of unscalable item score patterns using item response theory: An illustration with Harter's self-perception profile for children. *Journal of Personality Assessment*.
- Meijer, R. R., & Sijtsma, K. (1995). Detection of aberrant item score patterns: A review of recent developments. *Applied Measurement in Education*, *8*, 261–272.
- Meijer, R. R., & Sijtsma, K. (2001). Methodological review: Evaluating person fit. *Applied Psychological Measurement*, *25*, 107–135.
- Miller, J. D., Pilkonis, P. A., & Clifton, A. (2005). Self- and other-reports of traits from the five-factor model: Relations to personality disorders. *Journal of Personality Disorders*, *19*, 400–419.
- Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, *17*, 297–334.
- Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association*, *78*, 47–65.
- Muthén, L. K., & Muthén, B. O. (2006). Mplus: Statistical analysis with latent variables (Version 4.12) [Computer software]. Los Angeles: Author.

- Nering, M. L., & Meijer, R. R. (1998). A comparison of the person response function and the I_z person fit statistic. *Applied Psychological Measurement, 22*, 53–69.
- Oltmanns, T. F., & Turkheimer, E. (2006). Perceptions of self and others regarding pathological personality traits. In R. F. Krueger & J. Tackett (Eds.), *Personality and psychopathology: Building bridges* (pp. 71–111). New York: Guilford.
- Piedmont, R. L., McCrae, R. R., Riemann, R., & Angleitner, A. (2000). On the invalidity of validity scales: Evidence from self-reports and observer ratings in volunteer samples. *Journal of Personality and Social Psychology, 78*, 582–593.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
- Reise, S. P. (1995). Scoring method and detection of person misfit in a personality assessment context. *Applied Psychological Measurement, 19*, 213–229.
- Reise, S. P. (2000). Using multilevel logistic regression to evaluate person fit in IRT models. *Multivariate Behavioral Research, 35*, 543–568.
- Reise, S. P., & Flannery, P. W. (1996). Assessing person fit on measures of typical performance. *Applied Measurement in Education, 9*, 9–26.
- Reise, S. P., & Waller, N. G. (1993). Traitedness and the assessment of response pattern scalability. *Journal of Personality and Social Psychology, 65*, 143–151.
- Reise, S. P., & Waller, N. G. (2003). How many IRT parameters does it take to model psychopathology items? *Psychological Methods, 8*, 164–184.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461–464.
- Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*, 159–194.
- Sijtsma, K., & Meijer, R. R. (2001). The person response function as a tool in person-fit research. *Psychometrika, 66*, 191–208.
- Snijders, T., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage.
- Tellegen, A. (1982). *Brief manual of the Multidimensional Personality Questionnaire*. Unpublished manuscript.
- Tellegen, A. (1988). The analysis of consistency in personality assessment. *Journal of Personality, 56*, 621–663.
- Thissen, D. (1991). MULTILOG user's guide: Multiple categorical item analysis and test scoring using item response theory. [Computer software]. Chicago: Scientific Software International.
- Thissen, D., & Wainer, H. (Eds.). (2001). *Test scoring*. Mahwah, NJ: Erlbaum.
- Thomas, C., Turkheimer, E., & Oltmanns, T. F. (2003). Factorial structure of pathological personality traits as evaluated by peers. *Journal of Abnormal Psychology, 112*, 1–12.
- Trabin, T. E., & Weiss, D. J. (1983). The person response curve: Fit of individuals to item response theory models. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 83–108). New York: Academic Press.
- Watson, D., Hubbard, B., & Weise, D. (2000). Self–other agreement in personality and affectivity: The role of acquaintanceship, trait visibility, and assumed similarity. *Journal of Personality and Social Psychology, 78*, 546–558.
- Woods, C. M. (2006a). Ramsay-curve item response theory to detect and correct for nonnormal latent variables. *Psychological Methods, 11*, 253–270.
- Woods, C. M. (2006b). *RCLOG .v. 2: Software for item response theory parameter estimation with the latent population distribution represented using spline-based densities*. (Tech. Rep.). St. Louis, MO: Washington University in St. Louis.
- Woods, C. M. (2007). Ramsay-curve IRT for Likert-type data. *Applied Psychological Measurement, 31*, 195–212.
- Woods, C. M. (in press). Consequences of ignoring guessing when estimating the latent density in item response theory. *Applied Psychological Measurement*.
- Woods, C. M. (2008). Monte Carlo evaluation of two-level logistic regression for assessing person fit. *Multivariate Behavioral Research, 43*, 50–76.
- Woods, C. M., & Thissen, D. (2006). Item response theory with estimation of the latent population distribution using spline-based densities. *Psychometrika, 71*, 281–301.
- Zimowski, M., Muraki, E., Mislevy, R., & Bock, D. (2003). BILOG-MG 3. [Computer software]. Lincolnwood, IL: Scientific Software International.
- Zumbo, B. D., Pope, G. A., Watson, J. E., & Hubley, A. M. (1997). An empirical test of Roskam's conjecture about the interpretation of an ICC parameter in personality inventories. *Educational and Psychological Measurement, 57*, 963–969.

Received June 14, 2007

Revision received January 29, 2008

Accepted February 5, 2008 ■