# Fitting the Two-Parameter Model to Personality Data

**Steven P. Reise and Niels G. Waller**
**University of Minnesota**

The Multidimensional Personality Questionnaire (MPQ; Tellegen, 1982) was parameterized using the two-parameter logistic item response model. This entailed assessment of the suitability of personality data for item response analyses, including the assessment of dimensionality, monotonicity of item response, and data-model fit. The latter issue received special emphasis. Similarities and differences between maximum performance and typical performance data are discussed in relation to item response theory. Results suggest that the two-parameter model fits the MPQ data and that researchers engaged in the assessment of normal-range personality processes have much to gain from exploiting item response models. *Index terms: item fit, item response theory, Multidimensional Personality Questionnaire, personality measurement, two-parameter model.*

Within the family of item response models, the two-parameter logistic model (Birnbaum, 1968; Maxwell, 1959) is interesting for two reasons. First, in contrast to the more popular one- and three-parameter models, there is a dearth of research pertaining to its application. Second, researchers have yet to apply this model to what appears to be its most suitable application, namely, the measurement of multiple personality variables within a single inventory. Although item response theory (IRT) has been applied to selected personality scales (Bejar, 1977; Carter & Wilkinson, 1985; de Jong-Gierveld & Kamphuis, 1985; Sapinkopf, 1977), no study has applied the two-parameter logistic model to an entire personality inventory.

This study reports on the parameterization of the Multidimensional Personality Questionnaire (MPQ; Tellegen, 1982; Tellegen & Waller, in press) using the two-parameter IRT model. The MPQ was developed over a 15-year time span during which numerous rounds of item writing, data collection and analysis, and construct redefinition took place. As a result of this effort, the 11 MPQ scales are relatively unidimensional and independent. These properties make the MPQ an appropriate candidate for an IRT analysis.

A primary objective of this report is to present several methods for assessing the suitability of personality data for item response analyses. Special emphasis is placed on the issue of model-data fit in the context of typical performance (i.e., personality) data. Heretofore, most applications of IRT have focused on maximum performance data, as found in the assessment of ability or achievement. Personality assessment, on the other hand, is unique in its concern with typical or average performance; hence it is necessary to establish whether item response models may be legitimately applied to data of this type. This project is exploratory in that an attempt was made to search for and to describe problems germane to parameterizing small sets of typical performance items.

*APPLIED PSYCHOLOGICAL MEASUREMENT*
*Vol. 14, No. 1, March 1990, pp. 45–58*

*0146-6216/90/010045-14$1.95*

## Method

### The Multidimensional Personality Questionnaire

The MPQ is a 300-item personality inventory designed to assess a broad array of normal-range personality constructs. The instrument was developed "out of an attempt to clarify both the structure and content of the self-view domain" (Tellegen, 1982, p. 1). Eleven primary traits and three higher-order factors are measured by this instrument. The analyses reported in this paper were performed on the 284 mostly True-False items that comprise the 11 content scales. The 11 scale names and the number of items in each scale are given in Table 1.

### Calibration Samples

Two independent samples of 1,000 adults were used in the calibration runs. The 2,000 persons were randomly selected from the more than 6,000 persons in the Minnesota Twin Registry. The aggregate sample consisted of 1,127 females (mean age = 40.78, standard deviation = 9.36, range = 66 years) and 873 males (mean age = 43.09, standard deviation = 10.6, range = 67 years).

### Procedure

Items were calibrated using BILOG (Mislevy & Bock, 1986), which uses the marginal maximum likelihood algorithm (Bock, 1972; Bock & Aitkin, 1981; Bock & Lieberman, 1970; Mislevy & Bock, 1982; Mislevy & Stocking, 1989; Thissen, 1982) to estimate the item parameters of the two-parameter model (Birnbaum, 1968), which is expressed as

$$P|\theta = \frac{1}{1+\exp[-a(\theta-b)]}, \qquad (1)$$

where $\theta$ is the continuous $[-\infty,\infty]$ latent trait underlying test performance,
$P|\theta$ is the probability of a keyed item response conditional on $\theta$,
$a$ is the item discrimination, and
$b$ is the item difficulty.

The identification problem (Hambleton & Swaminathan, 1985, p. 126) was solved by setting the $\theta$ scale to have a mean of 0 with standard deviation of 1.0 in the calibration samples.

Item response models can be used to reflect item response behavior when the data meet several fundamental criteria. According to these criteria, item responses within each scale are determined by a single underlying trait, are locally independent, and conform to the model and therefore are monotone increasing functions of the latent trait (Hambleton & Swaminathan, 1985; Lord, 1980).

## Analyses and Results

### Number-Keyed Score Analyses

Table 1 presents descriptive statistics for the raw scores (number of items answered in the keyed direction) for the 11 MPQ scales. As the table shows, the 11 scales differ in their first and third moments. Although some scales (e.g., Absorption) have relatively Gaussian distributions, other scales (e.g., Alienation) are relatively skewed. Given these differences, it is important to investigate whether the shapes of the raw-score distributions were associated with model-data fit; this issue is considered below.

*Choice of model.* An initial concern with personality data is whether item responses can be adequately represented by the family of logistic ogives. If it is shown that the logistic ogive accurately portrays personality response behavior, a second concern is whether the two-parameter logistic model is the best model from the family of possible item-response models. Specifically, with personality data, the probabilities of a keyed response might not have lower asymptotes near 0 and/or upper asymptotes near 1.

To investigate these concerns, local difficulty (LD) vectors were computed for all 284 MPQ items using the aggregated sample of 2,000 persons. LD vectors represent the proportions of persons responding in the keyed direction conditional on having the same raw scale score (minus the item being investigated).

Table 1
Descriptive Statistics for 11 MPQ Scales

| Scale | No. of Items | Mean | Median | Variance | Kurtosis | Skewness |
|---|---|---|---|---|---|---|
| Well-Being (WB) | 24 | 18.56 | 20.00 | 24.07 | 1.24 | −1.26 |
| Social Potency (SP) | 26 | 9.46 | 9.00 | 42.91 | −.74 | .49 |
| Achievement (ACH) | 21 | 12.37 | 13.00 | 18.52 | −.52 | −.28 |
| Social Closeness (SC) | 22 | 14.95 | 16.00 | 22.81 | −.40 | −.58 |
| Stress Reaction (SR) | 26 | 10.87 | 10.00 | 45.47 | −.95 | .29 |
| Aggression (AGG) | 20 | 3.90 | 3.00 | 11.82 | 1.55 | 1.26 |
| Alienation (AL) | 20 | 2.39 | 1.00 | 11.61 | 5.37 | 2.21 |
| Control (CN) | 24 | 16.10 | 17.00 | 21.83 | −.23 | −.52 |
| Harm Avoidance (HA) | 28 | 20.92 | 22.00 | 29.38 | .64 | −.99 |
| Traditionalism (TRAD) | 27 | 18.91 | 20.00 | 26.61 | .16 | −.81 |
| Absorption (ABS) | 34 | 15.12 | 14.00 | 50.72 | −.49 | .29 |

At the raw-score level, the overwhelming majority of items conformed to the model's specifications (i.e., monotonicity, lower asymptotes of 0.0, upper asymptotes of 1.0). However, several items had endorsement probabilities greater than 0 for persons with very low (i.e., 0, 1, 2) raw scale scores, and a few items had endorsement probabilities less than 1 for high raw scale scores. Table 2 displays the LD vectors for several items characteristic of these observations.

The LD vectors, illustrated in the first three columns of Table 2, may reflect low item discrimination, extreme (high or low) item difficulty, and/or bounded raw scores. They may also indicate that a different latent trait model is more appropriate for these items. For example, some MPQ items might be more accurately modeled by the three-parameter (Lord & Novick, 1968) or four-parameter (Barton & Lord, 1981) models. However, as discussed below, most of the items that appeared to be three- or four-parameter functions at the raw-score level fit the two-parameter model quite well when the scores were determined on the $\theta$ metric. Figure 1 illustrates this point using Achievement item 36, whose LD vector is shown in Table 2.

*One-parameter versus two-parameter model.* While the LD vectors indicate that a three- or four-parameter model is not needed for the MPQ items, they do not provide the information necessary to choose between a one- or a two-parameter model. Two analyses were performed to investigate this issue. First, classical item discrimination estimates (point-biserial correlations) and difficulty estimates (proportion endorsed) were computed for each of the 284 items based on the aggregated sample of 2,000 persons. The MPQ items differ markedly in endorsement propensity (difficulty); although this is a desirable characteristic, it is known to affect the range of the item-test point-biserials (Lord & Novick, 1968). Therefore, item-scale biserial correlations, which are not confounded by item difficulty, were computed.

A substantial amount of variation existed between the item-scale biserials. This suggested that the one-parameter model was not appropriate. To investigate this issue further, standard errors of the biserial correlations based on the mean (within-scale) endorsement frequencies were computed. Ratios of the within-scale standard deviations of the item-scale biserials to the average within-scale biserial standard errors were then computed. These ratios indicated that more variation existed between the item-scale biserials than can be accounted for by sampling error. In all cases the standard deviations of the item-scale biserials were substantially larger than the average biserial standard errors. This result offered further support for the use of the two-parameter model with this inventory.

Table 2
Local Difficulty (LD) Vectors for Four MPQ Items

| Raw Score | Item 36[a] ACH LD | ACH N | Item 291[b] SC LD | SC N | Item 35[c] SP LD | SP N | Item 108[d] WB LD | WB N |
|---|---|---|---|---|---|---|---|---|
| 0 | .75 | 4 | .50 | 2 | .03 | 64 | .00 | 4 |
| 1 | .62 | 9 | .38 | 8 | .00 | 139 | .00 | 4 |
| 2 | .79 | 17 | .36 | 13 | .01 | 130 | .00 | 10 |
| 3 | .65 | 36 | .50 | 12 | .02 | 133 | .20 | 5 |
| 4 | .72 | 50 | .62 | 26 | .02 | 108 | .08 | 13 |
| 5 | .62 | 54 | .68 | 37 | .03 | 118 | .25 | 12 |
| 6 | .69 | 83 | .65 | 54 | .02 | 103 | .18 | 22 |
| 7 | .79 | 100 | .66 | 59 | .06 | 107 | .39 | 23 |
| 8 | .72 | 118 | .68 | 73 | .07 | 92 | .22 | 23 |
| 9 | .81 | 142 | .79 | 96 | .09 | 117 | .48 | 27 |
| 10 | .80 | 167 | .77 | 84 | .05 | 92 | .45 | 29 |
| 11 | .78 | 170 | .79 | 92 | .10 | 88 | .34 | 50 |
| 12 | .80 | 183 | .87 | 127 | .08 | 98 | .51 | 65 |
| 13 | .81 | 171 | .83 | 128 | .06 | 85 | .55 | 51 |
| 14 | .83 | 164 | .82 | 125 | .11 | 54 | .60 | 72 |
| 15 | .82 | 142 | .82 | 151 | .21 | 67 | .61 | 85 |
| 16 | .82 | 144 | .85 | 168 | .24 | 71 | .70 | 102 |
| 17 | .83 | 102 | .88 | 186 | .19 | 54 | .71 | 126 |
| 18 | .85 | 90 | .87 | 158 | .27 | 52 | .79 | 163 |
| 19 | .70 | 37 | .88 | 161 | .35 | 60 | .82 | 190 |
| 20 | .83 | 19 | .92 | 142 | .39 | 38 | .91 | 222 |
| 21 | | | .93 | 98 | .43 | 44 | .90 | 250 |
| 22 | | | | | .32 | 37 | .93 | 248 |
| 23 | | | | | .64 | 22 | .99 | 204 |
| 24 | | | | | .53 | 17 | | |
| 25 | | | | | .60 | 10 | | |
| $\chi^2$ (8 *df*) | 17.46 | | 15.31 | | 10.87 | | 12.30 | |

[a]Did not have lower asymptote of 0 or upper asymptote of 1.
[b]Did not have lower asymptote of 0.
[c]Did not have upper asymptote of 1.
[d]A typical two-parameter ogive function.

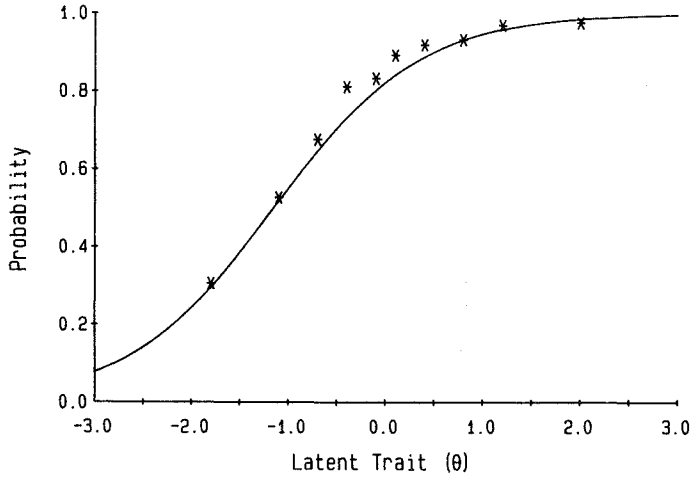
As a second analysis, 1,000 responses from each of the 11 MPQ scales were parameterized with one- and two-parameter models using BILOG (Mislevy & Bock, 1986). The BILOG goodness-of-fit statistics are reported in Table 3. The table clearly shows that the MPQ scales are more accurately described by a two-parameter model.

### Unidimensionality and Local Independence

The unidimensionality assumption was examined by comparing the ratio of the first to the second eigenvalue for each within-scale matrix of tetrachoric correlations. This ratio is an index of the strength of the first dimension of the data. Strictly speaking, the IRT model assumes that an examinee's position in the latent space can be accounted for completely by one latent dimension. This assumption, however, is rarely if ever met in practice (Hambleton & Swaminathan, 1985, p. 17). A more realistic requirement, therefore, is that the first dimension of a dataset account for a substantial proportion of the matrix variance.

Table 4 presents the eigenvalues from this anal-

**Figure 1**
IRF and Observed Response Probabilities for Achievement Item 36
($\chi^2 = 17.40$, $a = 1.36$, $b = -1.15$)




ysis for the 11 MPQ scales. These eigenvalues were obtained by performing an eigenstructure decomposition on the 11 tetrachoric correlation matrices using LISCOMP (Muthén, 1987). Because a conclusive test of the dimensionality assumption is not available (see Hattie, 1985), the data were examined to determine if one dominant trait accounted for the item intercorrelations (see Lord, 1980, pp. 20–21). The table shows that the assumption of a dominant dimension underlying each scale was well founded, because the first dimension accounted for considerably more variance than any other dimension for all scales.

The local-independence assumption was not directly examined because no satisfactory tests of this assumption currently exist (however, see Rosenbaum, 1984; Stout, 1987).

## Parameterizations

The 11 MPQ scales were parameterized twice (two samples of 1,000 persons) using BILOG (Mis-

Table 3
Goodness-of-Fit Summary From BILOG After Parameterizing the 11 MPQ Scales With Both a One- and Two-Parameter Logistic Model

| | One-Parameter | | | Two-Parameter | | |
|---|---|---|---|---|---|---|
| Scale | $\chi^2$ | *df* | *p* | $\chi^2$ | *df* | *p* |
| Well-Being | 203 | 142 | <.01 | 102 | 143 | .99 |
| Social Potency | 340 | 189 | <.01 | 173 | 207 | .95 |
| Achievement | 335 | 183 | <.01 | 204 | 162 | .01 |
| Social Closeness | 336 | 157 | <.01 | 153 | 165 | .72 |
| Stress Reaction | 509 | 217 | <.01 | 190 | 214 | .87 |
| Aggression | 217 | 108 | <.01 | 124 | 106 | .11 |
| Alienation | 122 | 72 | <.01 | 53 | 67 | .88 |
| Control | 361 | 187 | <.01 | 180 | 183 | .53 |
| Harm Avoidance | 293 | 217 | <.01 | 152 | 215 | .99 |
| Traditionalism | 404 | 217 | <.01 | 219 | 218 | .45 |
| Absorption | 312 | 280 | <.01 | 244 | 289 | .97 |

Table 4
Eigenvalues for 11 MPQ Scales Based on Tetrachoric Correlations ($N$ = 2,000)

| Eigenvalue | WB | SP | ACH | SC | SR | AGG | AL | CN | HA | TRAD | ABS |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 10.99 | 12.1 | 6.61 | 8.78 | 11.80 | 7.46 | 10.65 | 7.72 | 9.84 | 8.38 | 10.80 |
| 2 | 1.96 | 1.77 | 2.33 | 1.78 | 1.68 | 1.58 | 1.46 | 1.76 | 2.08 | 2.34 | 2.09 |
| 3 | 1.20 | 1.59 | 1.55 | 1.35 | 1.32 | 1.38 | 1.20 | 1.65 | 1.29 | 1.67 | 1.52 |
| 4 | .98 | 1.21 | 1.14 | 1.28 | 1.07 | 1.14 | .90 | 1.64 | 1.04 | 1.48 | 1.21 |
| 5 | .87 | .98 | 1.07 | 1.05 | .95 | .98 | .81 | 1.15 | 1.00 | 1.10 | 1.14 |
| 6 | .78 | .91 | .89 | .93 | .85 | .96 | .65 | 1.06 | .92 | 1.00 | 1.05 |
| 7 | .74 | .79 | .80 | .77 | .80 | .77 | .55 | .96 | .86 | .96 | 1.03 |
| 8 | .66 | .66 | .76 | .71 | .72 | .71 | .49 | .84 | .85 | .81 | .99 |
| 9 | .65 | .62 | .71 | .64 | .67 | .67 | .47 | .78 | .78 | .79 | .87 |
| 10 | .63 | .53 | .65 | .62 | .65 | .62 | .44 | .74 | .78 | .76 | .84 |
| 11 | .52 | .50 | .62 | .54 | .53 | .54 | .41 | .66 | .71 | .74 | .80 |
| 12 | .48 | .48 | .54 | .51 | .52 | .52 | .35 | .60 | .69 | .67 | .75 |
| 13 | .45 | .45 | .52 | .44 | .48 | .48 | .32 | .57 | .64 | .63 | .72 |
| 14 | .43 | .41 | .49 | .41 | .47 | .41 | .28 | .51 | .62 | .61 | .71 |
| 15 | .38 | .38 | .42 | .34 | .45 | .40 | .24 | .48 | .59 | .56 | .68 |
| 16 | .37 | .37 | .39 | .32 | .39 | .34 | .17 | .46 | .57 | .53 | .64 |
| 17 | .34 | .33 | .35 | .30 | .38 | .30 | .15 | .43 | .52 | .47 | .62 |
| 18 | .31 | .28 | .32 | .29 | .33 | .25 | .14 | .39 | .49 | .46 | .61 |
| 19 | .29 | .28 | .30 | .25 | .33 | .23 | .12 | .34 | .47 | .44 | .59 |
| 20 | .23 | .25 | .26 | .23 | .31 | .16 | .10 | .30 | .46 | .44 | .58 |
| 21 | .21 | .23 | .20 | .19 | .27 | | | .27 | .44 | .40 | .57 |
| 22 | .18 | .19 | | .16 | .23 | | | .26 | .40 | .35 | .52 |
| 23 | .13 | .19 | | | .23 | | | .20 | .38 | .32 | .50 |
| 24 | .10 | .14 | | | .19 | | | .13 | .35 | .30 | .47 |
| 25 | | .10 | | | .14 | | | | .32 | .28 | .44 |
| 26 | | .09 | | | .09 | | | | .29 | .22 | .42 |
| 27 | | | | | | | | | .28 | .16 | .41 |
| 28 | | | | | | | | | .24 | | .39 |
| 29 | | | | | | | | | | | .37 |
| 30 | | | | | | | | | | | .36 |
| 31 | | | | | | | | | | | .35 |
| 32 | | | | | | | | | | | .34 |
| 33 | | | | | | | | | | | .33 |
| 34 | | | | | | | | | | | .29 |

levy & Bock, 1986). All BILOG default values were used, with the exception that an empirical distribution of examinee latent trait scores was computed after each iteration. This option was implemented because scores for several of the MPQ scales (e.g., Alienation) were not normally distributed.

*Item parameters across calibrations.* An item calibration performed twice using different samples will not produce item parameters on the same scale. However, if the model holds, the parameters will be highly related because in theory IRT item parameters are invariant up to a linear transformation within a defined population (Hambleton & Swaminathan, 1985). Table 5 presents descriptive statistics for these parameters and the correlations between the parameter estimates for the two calibrations. Only the descriptive statistics for the first sample are shown because the means and standard deviations were essentially equivalent across samples.

As seen in Table 5, the correlations between parameter estimates across calibrations are uniformly high, especially for the item difficulty estimates (median $r = .88$). The item discriminations correlated above .80, which was considered satisfactory given that (1) any likelihood-based program will have problems estimating this parameter (Thissen & Wainer, 1982), and (2) by design, the range of discriminations was somewhat restricted as compared to the range of item difficulties.

*Item position effects.* A concern with personality data is that the item discrimination parameters may be correlated with the location of the items in the test booklet (Knowles, 1988). Strong item discrimination-position correlations may indicate that the local-independence assumption is violated. Several factors unique to personality data have been shown to contribute to item discrimination-position correlations. Knowles (1988), for example, has shown that for single-trait personality scales, item-scale correlations generally increase in later sections of an inventory.

Knowles' findings were based on single-construct measures, and therefore are not likely to apply to inventories such as the MPQ that do not present items by scale (i.e., the items within a scale are interspersed throughout the inventory). However, as a precautionary step the independence of item discrimination and item location was investigated. Correlations between the item discrimination, squared item discrimination, and booklet item numbers across the 284 items were computed. The results of this analysis suggest that these variables are not linearly related to item position. The average correlations across scales were $r = .11$ for discrimination and $r = .11$ for squared discrimination.

## Model-Data Fit

Having generated two sets of parameter estimates for each scale, an analysis was conducted to determine which set of parameters best represented the entire sample. Item parameters derived from the first sample were used to compute maximum likelihood $\theta$ estimates for persons in the second sample. Response vectors of persons answering all or none of the items in the keyed direction received boundary scores of 4.00 and $-4.00$, respectively. Chi-square tests of fit for the second sample were then computed using parameters from the first sample. This procedure was repeated with the param-

Table 5
Descriptive Statistics for the Item Parameters Based on the First Sample of 1,000 Responses and Correlations Between Item Parameters From the Two Calibration Samples

| | *a* | | *b* | | Correlations | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | Across Samples | | Within Sample | |
| | | | | | | | 1 | 2 |
| Scale | Mean | SD | Mean | SD | *a* | *b* | *ab* | *ab* |
| Well-Being | 1.50 | .43 | 1.20 | .59 | .91 | .96 | .16 | .03 |
| Social Potency | 1.56 | .39 | .62 | .64 | .88 | .99 | .36 | .43 |
| Achievement | 1.08 | .46 | .56 | 1.36 | .93 | .96 | .20 | .13 |
| Social Closeness | 1.36 | .44 | .99 | 1.07 | .93 | .98 | .42 | .51 |
| Stress Reaction | 1.47 | .36 | .33 | .49 | .86 | .97 | .12 | .25 |
| Aggression | 1.29 | .43 | 1.71 | .81 | .90 | .95 | .37 | .46 |
| Alienation | 1.80 | .47 | 1.75 | .40 | .86 | .93 | .35 | .04 |
| Control | 1.16 | .43 | .87 | .83 | .87 | .97 | .34 | .05 |
| Harm Avoidance | 1.19 | .29 | 1.32 | .51 | .80 | .86 | .04 | .23 |
| Traditionalism | 1.10 | .42 | 1.20 | .89 | .91 | .93 | .09 | .26 |
| Absorption | 1.12 | .19 | .29 | .89 | .81 | .99 | .03 | .17 |

eters from the second sample and the response vectors from the first sample. For each scale, the set of parameters that demonstrated the best $\chi^2$ fit-to-model result on the independent sample was selected as the final parameter set.

In the assessment of model-data fit a modified version of Bock's (1972) $\chi^2$ was employed. The modification was to construct 10 $\theta$ intervals containing an equal number of examinees per interval. This statistic has the following form:

$$\chi^2 = \sum_{j=1}^{10} \frac{N_j(O_{ij}-E_{ij})^2}{E_{ij}(1-E_{ij})} \quad , \tag{2}$$

where $i$ is the item,
$j$ is the interval created by grouping persons on the basis of their $\theta$ estimates (10 intervals of equal $N$ were used in all analyses),
$N_j$ is the number of persons with $\theta$ estimates falling in the $j$th interval,
$O_{ij}$ is the observed proportion of keyed responses on item $i$ for interval $j$, and
$E_{ij}$ is the expected proportion of keyed responses on item $i$ for interval $j$ based on computing the item response function (IRF) evaluated at the median $\theta$ estimate within the interval.

Bock's $\chi^2$ statistic is one of many $\chi^2$ statistics used to assess model-data fit in IRT. McKinley and Mills (1985) have shown that most common $\chi^2$ tests (e.g., Wright & Mead, 1977; Yen, 1981) produce comparable results with normally distributed data. In the present study Bock's $\chi^2$ was used because this index has been shown to yield the smallest Type I error rate of the commonly employed $\chi^2$ fit indices (McKinley & Mills, 1985).

Bock's $\chi^2$ statistic may be used to assess model-data fit at two levels of analysis: (1) at the item level, and (2) at the scale level by summing the individual item $\chi^2$s within a scale. These fit indices should not be interpreted as conclusive evidence of adequate model fit, as they are extremely sensitive to several factors including the construction of the intervals, the number of intervals created, and the sample size. But even within these limitations Bock's $\chi^2$ can be used legitimately to suggest which set of parameters is more robust across samples and to tentatively identify misfitting items.

In the present dataset, persons whose responses were in either the all keyed or all non-keyed direction could be retained with little effect on the $\chi^2$ for all scales except Alienation. In this scale, 760 persons with all zero response vectors were eliminated for the analyses described below. The elimination of these response vectors resulted in improved fit.

After determining which set of parameters exhibited greater robustness across samples, maximum likelihood score estimates for each scale were computed for all 2,000 persons. $\chi^2$ values were then computed and a residual table and residual plots (i.e., the plot of the theoretical and empirical IRFs) were generated. The residual plots depict the observed proportion of keyed responses within each of the 10 $\theta$ intervals minus the predicted proportions at the interval median $\theta$ estimate. The residual table is a 284 (items) $\times$ 10 ($\theta$ intervals per item) matrix consisting of the observed minus the expected proportions of persons answering the item in the keyed direction. Further discussion of residual analysis can be found in Hambleton and Swaminathan (1985, pp.184–193) or Kingston and Dorans (1985).

The $\chi^2$ fit-to-model ($N$ = 2,000) results are summarized in Table 6. At the item level, each $\chi^2$ is associated with 8 degrees of freedom ($df$ = 10 $\theta$ intervals minus 2 parameters). The $df$ associated with a given scale is computed by summing the item $df$ within that scale. With a sample size of 2,000 there was no expectation that the item- or scale-level $\chi^2$ would be nonsignificant at $\alpha$ = .05 (indicating fit to the model). A high value of $\chi^2$ for the present study was operationally defined as $\chi^2 > 26.13$, which is the $\alpha$ = .001 level of significance with 8 $df$. Thirty-six items meeting this criterion were identified. Of the 36 items tentatively identified as misfitting, 25 were from the Alienation, Aggression, and Harm Avoidance scales.

At the scale level, the ratio of $\chi^2$ to $df$ was computed as an index of scale fit. These ratios are shown in Table 6. Alienation, Aggression, and Harm Avoidance clearly have high ratios in comparison to the other scales. After identifying potentially

Table 6
Item and Scale $\chi^2$ Fit Values for the MPQ Under a Two-Parameter Model ($N$ = 2,000)

| Item | MPQ Scale | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| No. | WB | SP | ACH | SC | SR | AGG | AL | CN | HA | TRAD | ABS |
| 1 | 11.5 | 13.7 | 21.1 | 8.2 | 5.0 | 38.6* | 46.9* | 18.3 | 50.4* | 11.6 | 9.6 |
| 2 | 21.3 | 12.7 | 8.3 | 19.4 | 9.3 | 26.4* | 17.4 | 13.0 | 10.5 | 14.3 | 11.4 |
| 3 | 4.8 | 30.6* | 17.4 | 13.9 | 10.9 | 21.4 | 60.5* | 20.0 | 10.4 | 19.7 | 7.2 |
| 4 | 12.2 | 10.8 | 9.1 | 26.6* | 6.6 | 47.6* | 24.2 | 19.6 | 5.8 | 12.1 | 10.6 |
| 5 | 30.5* | 10.2 | 12.5 | 7.0 | 15.1 | 11.5 | 20.5 | 13.8 | 21.0 | 24.4 | 10.1 |
| 6 | 7.9 | 13.1 | 8.3 | 7.9 | 8.4 | 19.8 | 32.8* | 29.3* | 16.9 | 18.3 | 4.1 |
| 7 | 17.5 | 5.9 | 12.8 | 9.9 | 8.4 | 10.6 | 79.5* | 13.6 | 17.4 | 25.8 | 10.1 |
| 8 | 11.2 | 8.8 | 13.1 | 14.1 | 14.3 | 141.8* | 20.8 | 28.7* | 34.2* | 13.4 | 16.2 |
| 9 | 12.3 | 20.9 | 11.5 | 23.9 | 16.6 | 28.6* | 36.3* | 13.7 | 8.8 | 9.7 | 8.4 |
| 10 | 12.0 | 18.7 | 6.7 | 17.2 | 16.6 | 49.7* | 11.6 | 9.3 | 20.4 | 15.4 | 7.0 |
| 11 | 21.5 | 13.6 | 15.7 | 8.1 | 4.9 | 25.5 | 23.6 | 21.3 | 28.8* | 10.9 | 7.2 |
| 12 | 14.7 | 9.4 | 16.9 | 15.9 | 14.8 | 4.3 | 72.5* | 7.2 | 20.6 | 2.3 | 6.8 |
| 13 | 10.4 | 13.4 | 20.7 | 7.2 | 8.0 | 21.1 | 37.2* | 17.9 | 24.8 | 6.8 | 15.5 |
| 14 | 20.5 | 32.5* | 16.2 | 15.5 | 17.6 | 45.9* | 54.6* | 15.1 | 14.4 | 5.3 | 16.1 |
| 15 | 9.2 | 25.0 | 5.7 | 6.5 | 12.4 | 5.4 | 13.5 | 26.7* | 41.3* | 13.6 | 4.8 |
| 16 | 16.9 | 5.1 | 10.5 | 10.4 | 14.0 | 7.5 | 9.9 | 15.2 | 19.1 | 25.2 | 12.1 |
| 17 | 16.1 | 17.2 | 7.0 | 23.5 | 13.9 | 23.6 | 89.5* | 5.7 | 19.5 | 26.1 | 9.6 |
| 18 | 20.0 | 11.2 | 11.2 | 13.5 | 11.6 | 35.0* | 5.3 | 8.1 | 24.3 | 15.3 | 18.6 |
| 19 | 10.0 | 20.2 | 9.5 | 27.7* | 11.2 | 15.2 | 9.9 | 7.9 | 32.0* | 10.7 | 7.2 |
| 20 | 9.9 | 14.1 | 24.8 | 17.7 | 17.0 | 17.0 | 32.2* | 36.7* | 45.8* | 28.0* | 11.0 |
| 21 | 9.3 | 20.1 | 15.9 | 24.3 | 13.6 | | | 18.0 | 7.3 | 24.3 | 18.1 |
| 22 | 7.8 | 13.9 | | 15.3 | 7.8 | | | 4.6 | 31.0* | 15.0 | 20.8 |
| 23 | 9.4 | 12.1 | | | 22.4 | | | 12.4 | 16.0 | 21.8 | 8.9 |
| 24 | 15.0 | 14.2 | | | 22.3 | | | 5.6 | 16.1 | 10.3 | 14.4 |
| 25 | | 7.8 | | | 11.8 | | | | 30.9* | 10.4 | 12.7 |
| 26 | | 14.6 | | | 13.1 | | | | 10.8 | 10.2 | 10.2 |
| 27 | | | | | | | | | 19.9 | 5.0 | 18.5 |
| 28 | | | | | | | | | 6.7 | | 11.6 |
| 29 | | | | | | | | | | | 6.1 |
| 30 | | | | | | | | | | | 9.7 |
| 31 | | | | | | | | | | | 16.6 |
| 32 | | | | | | | | | | | 8.2 |
| 33 | | | | | | | | | | | 10.6 |
| 34 | | | | | | | | | | | 5.8 |
| Sum | 332 | 391 | 276 | 334 | 329 | 597 | 699 | 383 | 606 | 407 | 377 |
| $df$ | 192 | 208 | 168 | 176 | 208 | 160 | 160 | 192 | 224 | 216 | 272 |
| $\chi^2/df$ | 1.7 | 1.8 | 1.6 | 1.8 | 1.5 | 3.7 | 4.3 | 1.9 | 2.7 | 1.8 | 1.3 |

*$\chi^2$ greater than 26.13.

troublesome items, the item content, factor loadings (from an unweighted least-squares principal factor analysis on the tetrachoric correlation matrices), LD vectors, residual table, and residual plots were examined to determine possible causes of misfit.

If an MPQ item had a $\chi^2$ fit value less than 26.13, the theoretical and empirical IRFs within a given interval of $\theta$ (i.e., the cells in the residual table) tended to differ by no more than .05 on the probability metric. This observation, combined with an inspection of the residual plots, suggests that for

the majority of MPQ items, responses conformed to the logistic model. Figure 2a is a typical residual plot for an item with a $\chi^2$ value less than 26.13.

Of the 36 items tentatively identified as misfitting (denoted by asterisks in Table 6), 23 had $\chi^2$ values less than 40.0. Inspection of the residual table for these items revealed that for items with $\chi^2$ between 26.13 and 40.0 there were typically only one, or at most two, cells with deviant residuals (i.e., discrepancy more than .05). These deviant cells were inflating the $\chi^2$ values for these 23 items. The residual plots shown in Figures 2b and 2c exemplify how deviant responses within a single $\theta$ interval may inflate the $\chi^2$ statistic even though the model may fit over a broad trait range. This result suggests that for the majority of persons taking the MPQ, the item parameters for the items with $\chi^2$ between 26.13 and 40.0 adequately portray the observed response behaviors and would be acceptable for future work with this inventory (see Waller & Reise, 1989).

The remaining 13 items with $\chi^2$ values above 40.0 were from the Alienation (6), Aggression (4), and Harm Avoidance (3) scales. After inspecting both the classical and IRT item parameters, it became clear that the 13 items with the highest $\chi^2$ values were relatively easy items embedded in difficult scales, or relatively difficult items embedded in easy scales. For example, the three worst-fitting Harm Avoidance items (see Table 6) have the three lowest classical difficulties. Conversely, the worst-fitting Aggression item has the highest classical difficulty (i.e., this item has a relatively high endorsement frequency for this scale). Evidently, within these samples there are not enough persons with trait values in the ranges for which these items are most discriminating. Consequently, the estimation program was not provided with sufficient information and thus the parameters were poorly estimated.

Inspection of the residual plots for these 13 items confirmed this impression. Figure 2d exemplifies a residual plot for an item with a $\chi^2$ above 40.0. As can be seen in this figure, the IRF needs to be pulled to the right and made steeper. However, even with the relatively high $\chi^2$ value for this item (45.86) the plot shows that the degree of misfit is not great. This observation shows the necessity of using IRF plots in conjunction with point estimates of fit such as the $\chi^2$.

The LD vector analyses suggested that at the raw-score level the item responses are monotonic increasing functions. Also, the ranked eigenvalues of the tetrachoric item correlations suggested that a single dominant dimension underlies each scale. These observations implied that parameter estimation, not faulty item construction, caused the high $\chi^2$ values. In terms of item content and item factor loadings, there was no apparent difference between the misfitting items and the fitting items. However, as mentioned above, there is good reason to infer that the shape of the raw-score scale distribution is related to eventual fit of the two-parameter model, as estimated by BILOG. The more skewed the distribution (e.g., Alienation, Aggression), the more handicapped the estimation program will be, especially when estimating parameters for items that discriminate best in the tails of the distribution.

For a final analysis of fit, the residual table was partitioned by scale and examined. If the model fits, the within-scale row sums (the residuals across trait levels for each item within a given scale) and column sums (the residuals across items in a scale as a function of trait level) will exhibit a normal distribution centered at 0. Rows with a preponderance of positive or negative residuals indicate that the difficulty parameter has been poorly estimated. In the present dataset, there was no evidence that this occurred except for the 13 worst-fitting items previously discussed.

Column residual sums divided by the number of scale items indicate whether the IRFs are biased within specific sectors of the latent trait. Table 7 displays the results of summing the column residuals across items for each of the 11 MPQ scales. The table reveals no overwhelming evidence that the curves are consistently biased as a function of trait level.

## Conclusions

The results of this investigation suggest that the

**Figure 2**
IRFs and Observed Response Probabilities

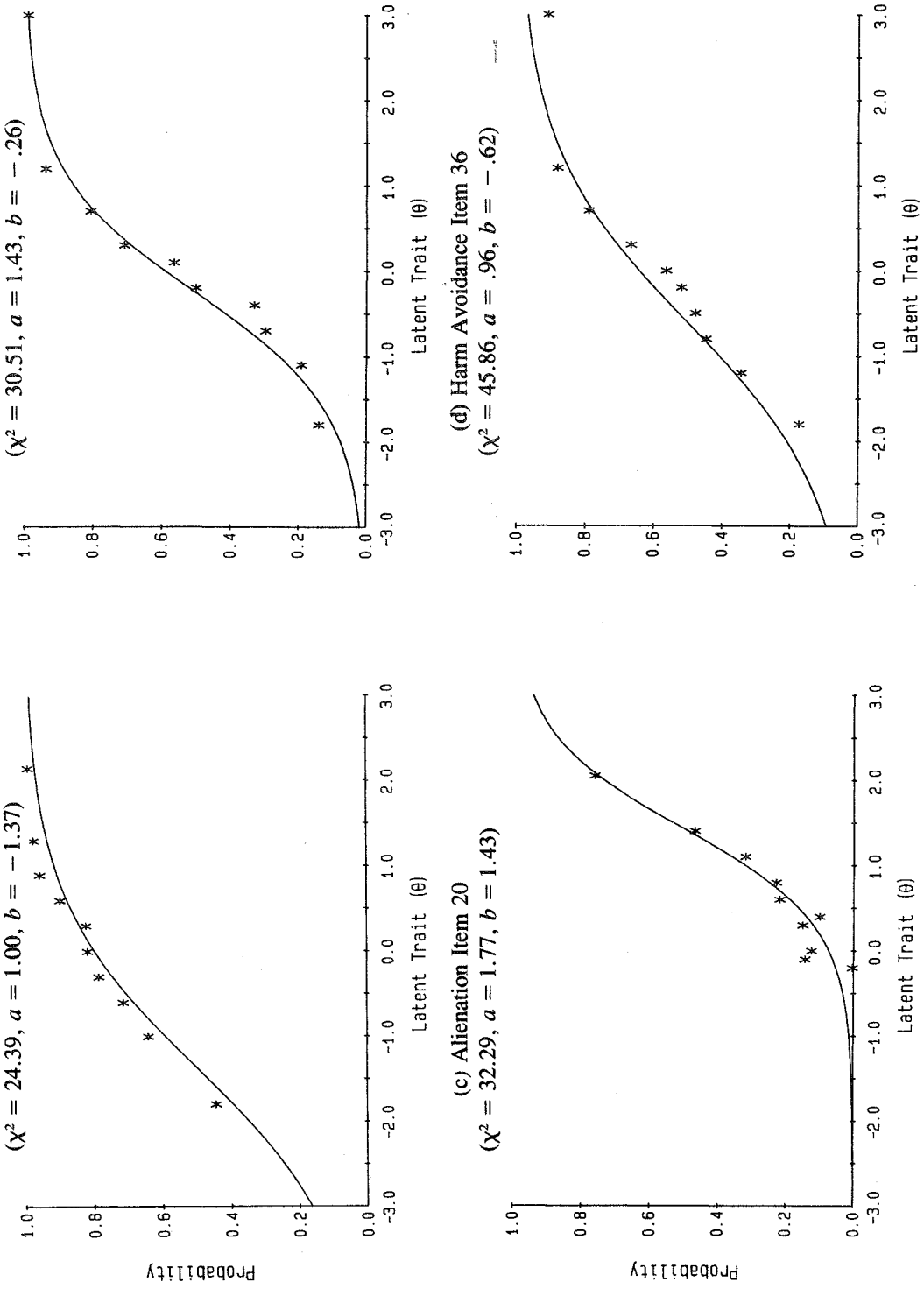


Table 7
Median θ Estimate and Residual Column Total as a Function of Trait Level

| Scale and Statistic | Grouped θ Interval 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Well-Being | | | | | | | | | | |
| Mdn θ estimate | −1.8 | −1.1 | −.7 | −.4 | −.2 | −.1 | .3 | .7 | 1.2 | 4.0 |
| Residual | .5 | .3 | .4 | .3 | 0.0 | .6 | .2 | 0.0 | 0.0 | 0.0 |
| Social Closeness | | | | | | | | | | |
| Mdn θ estimate | −1.7 | −1.1 | −.7 | −.4 | −.1 | .1 | .4 | .7 | 1.1 | 2.2 |
| Residual | .4 | .3 | .2 | .1 | .1 | .1 | 0.0 | 0.0 | .1 | .1 |
| Social Potency | | | | | | | | | | |
| Mdn θ estimate | −1.8 | −1.2 | −.7 | −.4 | −.1 | .1 | .4 | .7 | 1.1 | 1.9 |
| Residual | .2 | .2 | 0.0 | .2 | 0.0 | .3 | 0.0 | 0.0 | .3 | .1 |
| Achievement | | | | | | | | | | |
| Mdn θ estimate | −1.8 | −1.1 | −.7 | −.4 | −.1 | .1 | .4 | .8 | 1.2 | 2.2 |
| Residual | .4 | .2 | .1 | .1 | .1 | .2 | 0.0 | .2 | 0.0 | .1 |
| Stress Reaction | | | | | | | | | | |
| Mdn θ estimate | −2.1 | −1.2 | −.7 | −.4 | −.1 | .1 | .4 | .7 | 1.1 | 1.8 |
| Residual | 0.0 | 0.0 | .2 | .1 | 0.0 | .3 | .3 | .4 | .1 | .1 |
| Aggression | | | | | | | | | | |
| Mdn θ estimate | −4.0 | −1.3 | −1.0 | −.4 | −.1 | .1 | .4 | .8 | 1.1 | 1.8 |
| Residual | .1 | 0.0 | .1 | 0.0 | .1 | .3 | .2 | .2 | .1 | .8 |
| Alienation | | | | | | | | | | |
| Mdn θ estimate | −.2 | −.1 | 0.0 | .3 | .4 | .6 | .8 | 1.1 | 1.4 | 2.0 |
| Residual | .2 | 0.0 | .1 | .2 | .1 | .3 | .2 | .1 | .1 | .2 |
| Control | | | | | | | | | | |
| Mdn θ estimate | −1.7 | −1.1 | −.7 | −.4 | −.2 | −.1 | .3 | .7 | 1.2 | 2.3 |
| Residual | .5 | .1 | .1 | .2 | 0.0 | .2 | .2 | .1 | .2 | 0.0 |
| Traditionalism | | | | | | | | | | |
| Mdn θ estimate | −1.8 | −1.0 | −.6 | −.3 | 0.0 | .3 | .6 | .9 | 1.3 | 2.1 |
| Residual | .4 | 0.0 | .1 | .1 | 0.0 | 0.0 | .2 | 0.0 | 0.0 | .2 |
| Harm Avoidance | | | | | | | | | | |
| Mdn θ estimate | −1.8 | −1.2 | −.8 | −.5 | −.2 | 0.0 | .3 | .7 | 1.2 | 3.0 |
| Residual | .9 | .3 | .1 | 0.0 | 0.0 | .1 | .1 | 0.0 | .1 | .2 |
| Absorption | | | | | | | | | | |
| Mdn θ estimate | −1.8 | −1.0 | −.6 | −.4 | −.1 | .1 | .4 | .7 | 1.1 | 1.9 |
| Residual | .2 | .2 | .3 | .1 | .3 | 0.0 | .1 | .2 | .1 | .5 |

two-parameter logistic model fits the MPQ and that researchers engaged in the assessment of normal-range personality processes have much to gain from exploiting item response models. The present analysis revealed that the overwhelming majority of MPQ item responses conformed to the critical and necessary assumptions underlying the two-parameter logistic latent trait model. Although the findings cast a favorable light on the future use of IRT in the domain of normal-range personality assessment, several issues remain to be addressed concerning the applicability of IRT in the personality domain.

A primary question is whether parameters for items that do not fit the model should be reestimated, retained as is, or discarded. In the present study, items with $\chi^2$ below 26.13 were found to fit the model well, items with $\chi^2$ between 26.13 and 40.0 provided reasonable approximations to the observed response probabilities, and items with $\chi^2$ above 40.0 were found to be slightly problematic. However, future studies employing more hetero-

geneous samples on the Alienation, Aggression, and Harm Avoidance scales may provide data more amenable to obtaining satisfactory parameter estimates for these problematic items (see Stocking, 1988).

In personality scale research, a promising pursuit is to increase the range of item difficulties and reduce the corresponding peakedness of the test information curves. Peaked information curves are a common problem with classically derived scales (McBride, 1976). But for certain personality traits, unlike cognitive traits, it may prove difficult to write items that are discriminating in certain ranges of the latent trait. For example, it is difficult to conceive of Alienation items that discriminate maximally in a very low trait range.

Some personality traits may have an inherently quasi-categorical rather than a full-range continuum structure (Gangestad & Snyder, 1985). Consequently, for various personality traits, further developments in the application of IRT—such as item banking or adaptive testing (Waller & Reise, 1989)—may not be easily achieved without a Herculean item writing effort.

## References

Barton, M. A., & Lord, F. M. (1981). *An upper asymptote for the three-parameter logistic item-response model* (Research Bulletin 81-20). Princeton NJ: Educational Testing Service.

Bejar, I. I. (1977). An application of the continuous response level model to personality measurement. *Applied Psychological Measurement, 1,* 509–521.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 397–472). Reading MA: Addison-Wesley.

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

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika, 46,* 443–459.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for *n* dichotomously scored items. *Psychometrika, 35,* 179–197.

Carter, J. E., & Wilkinson, L. (1985). A latent trait analysis of the MMPI. *Multivariate Behavioral Research, 19,* 385–407.

de Jong-Gierveld, J., & Kamphuis, F. (1985). The development of a Rasch-type loneliness scale. *Applied Psychological Measurement, 9,* 289–299.

Gangestad, S., & Snyder, M. (1985). To curve nature at its joints: On the existence of discrete classes in personality. *Psychological Review, 92,* 317–349.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9,* 139–164.

Kingston, N. M., & Dorans, N. J. (1985). The analyses of item-ability regressions: An exploratory IRT model fit tool. *Applied Psychological Measurement, 9,* 281–288.

Knowles, E. S. (1988). Item context effects in personality scales: Measuring changes the measure. *Journal of Personality and Social Psychology, 55,* 312–320.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale NJ: Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading MA: Addison-Wesley.

Maxwell, A. E. (1959). Maximum likelihood estimates of item parameters using the logistic function. *Psychometrika, 24,* 221–227.

McBride, J. R. (1976). Bandwidth, fidelity and adaptive tests. In T. J. McConnell, Jr. (Ed.), *CAT/C21975: The second conference on computer assisted test construction*. Atlanta GA: Atlanta Public Schools.

McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement, 9,* 49–57.

Mislevy, R. J., & Bock, R. D. (1982). Implementation of the EM algorithm in the estimation of item parameters: The BILOG computer program. In D. J. Weiss (Ed.), *Proceedings of the 1982 Item Response Theory and Computerized Adaptive Testing Conference* (pp. 189–202). Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

Mislevy, R. J., & Bock, R. D. (1986). *BILOG I maximum likelihood item analysis and test scoring with binary logistic models* [Computer program]. Mooresville IN: Scientific Software Inc.

Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement, 13,* 57–76.

Muthén, B. (1987). *LISCOMP: Analysis of linear structural equations with a comprehensive measurement model* [Computer program]. Mooresville IN: Scientific Software Inc.

Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika, 49,* 425–436.

Sapinkopf, R. C. (1977). *A computer adaptive testing approach to the measurement of personality variables*. Unpublished doctoral dissertation, University of Maryland.

Stocking, M. L. (1988). *Specifying optimum examinees for item parameter estimation in item response theory* (Report No. RR-88-57-ONR). Princeton NJ: Educational Testing Service.

Stout, W. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika, 52,* 589–618.

Tellegen, A. (1982). *A brief manual for the Multidimensional Personality Questionnaire*. Unpublished manuscript, University of Minnesota.

Tellegen, A., & Waller, N. G. (in press). Exploring personality through test construction: Development of the Multidimensional Personality Questionnaire. In S. R. Briggs & J. M. Cheek (Eds.), *Personality measures: Development and evaluation* (Vol. 1). Greenwich CT: JAI Press.

Thissen, D. M. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika, 47,* 175–186.

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika, 47,* 397–412.

Waller, N. G., & Reise, S. P. (1989). Computerized adaptive personality assessment: An illustration with the Absorption scale. *Journal of Personality and Social Psychology, 57,* 1051–1058.

Wright, B. D., & Mead, R. J. (1977). *BICAL: Calibrating items and scales with the Rasch model* (Research Memorandum No. 23). Chicago IL: University of Chicago, Statistical Laboratory, Department of Education.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5,* 245–262.

## Acknowledgments

*The authors thank Yossi Ben Porath, Elana Broch, Auke Tellegen, and the Editor for their helpful comments on an earlier version of this paper. Special thanks are also due to David T. Lykken for allowing access to the Minnesota Twin Registry. This work was supported in part by grant MH37860 from the National Institute of Mental Health.*

## Author's Address

Send requests for reprints or further information to Steven Reise, 658 Elliott Hall, Department of Psychology, University of Minnesota, Minneapolis MN 55455, U.S.A.