Scoring Method and the Detection of Person Misfit in a Personality Assessment Context

Steven P. Reise

University of California, Riverside

The purpose of this research was to explore psychometric issues pertinent to the application of an IRT-based person-fit (response aberrancy) detection statistic in the personality measurement domain. Monte carlo data analyses were conducted to address issues regarding the l_z person-fit statistic. The major issues explored were characteristics of the null distribution of l_z and its power to identify nonfitting response patterns under different scoring strategies. There were two main results. First, the l_z index null distribution was not well standardized when item parameters of personality scales were used; the l_z null distribution variance was significantly less than the hypothesized value of 1.0 under several conditions. Second, the power of l_z to detect response misfit was affected by the scoring method. Detection power was optimal when a biweight estimator of θ was used. Recommendations are made regarding proper implementation of person-fit statistics in personality measurement. *Index terms: appropriateness measurement, item response theory, l_z statistic, person fit, personality assessment, response aberrancy, scoring methods, two-parameter model.*

In the context of personality trait measurement, there is little research directly pertaining to the psychometric issues involved in the application of person-fit (response aberrancy) statistics based on item response theory (IRT). Hence, this study addressed several basic quantitative issues pertaining to the application of the l_z person-fit statistic (Drasgow, Levine, & Williams, 1985) to personality trait measurement. In particular, two studies were conducted to address the following questions:

1. How is l_z distributed in a personality measurement context?
2. How powerful is l_z in detecting non-model-based responding (i.e., nonfitting response vectors)?
3. How is the power of l_z to detect nonfitting responses affected by the scoring method?

l_z (Drasgow et al., 1985) was the only person-fit statistic investigated. Although many potential person-fit indexes exist, l_z was selected for study for several reasons. First, l_z has generated a substantial amount of research in the ability measurement domain (e.g., Birenbaum, 1986; Drasgow, Levine, & McLaughlin, 1987; Drasgow et al., 1985). Also, in research that compared person-fit indexes (e.g., Gafni, 1988; Harnisch & Tatsuoka, 1983) this index has performed well—there were no indexes that consistently outperformed l_z.

Although l_z has functioned well in the ability assessment context, this research literature has certain characteristics. In particular, two properties of l_z have received the most empirical attention: (1) its conditional standardization on trait level (θ), and (2) its power to detect nonfit. However, in addressing these issues, researchers (e.g., Drasgow et al., 1987) have computed l_z on a single, long ability test such as the SAT or GRE Verbal. There has not been any research suggesting how different test lengths and different item parameter distributions affect l_z's standardization and power in the context of personality data that is fit to a two-parameter IRT model.

Previous research on the power of l_z to detect nonfit (e.g., Gafni, 1988; Reise & Due, 1991; Schmitt, Cortina, & Whitney, 1993) has relied extensively on using either true θ (in simulated data) or maximum likelihood (ML) estimates of $\hat{\theta}$ (in real data) for computing l_z. However, there are many methods for esti-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 19, No. 3, September 1995, pp. 213–229

0146-6216/95/030213-17$2.10

mating θ other than ML, and some of these methods may affect l_z's ability to detect misfitting responses. The standardization implicit in l_z is based on true person and item parameters. Hence, l_z is optimal when attempting to detect misfitting response patterns that are unlikely given an examinee's θ level. In applied situations, $\hat{\theta}$ must be used to compute l_z. How the scoring method affects the power to detect person nonfit is an unexplored issue.

To address these issues, two studies were conducted. The studies explored the statistical properties of l_z when computed for simulated item responses to several personality measurement scales. In Study 1, simulated item responses based on personality scale item parameter estimates were used to address l_z distributional issues. Simulated response vectors also were used in Study 2 to address issues of power and the effects of different scoring methods.

Study 1: The Null Distribution of l_z

Study 1 examined the distribution of l_z to determine whether l_z is well standardized when items with parameters characteristic of personality tests were used. The specific question raised in Study 1 was whether l_z had a θ conditional null distribution that is consistently standard normal across a reasonable range of θ levels.

Researchers (Drasgow et al., 1985) using l_z in ability assessment contexts have proposed that l_z is distributed approximately as standard normal, conditional on θ, when the item response patterns fit the model. Previous research has been based on computing l_z in the context of relatively long ability test data (e.g., SAT Verbal with 95 items) that had been fit to the three-parameter model. The personality scales used in this research, which ranged from 24 to 34 items in length, are shorter compared to the scales typically used in person-fit research and were fit to the two-parameter logistic model (2PLM).

Method

Personality Trait Measures

The Multidimensional Personality Questionnaire (MPQ; Tellegen, 1982) was used as the basis for all analyses. The MPQ consists of 284 items to which respondents answer *true* or *false*. All scales were developed by factor analytic methods detailed in Tellegen and Waller (in press). In particular, four trait scales from the MPQ were selected: Control (CO, 24 items), Harm Avoidance (HA, 28 items), Traditionalism (TR, 27 items), and Absorption (AB, 34 items). These measures were selected because they contain the most items in the MPQ inventory, and previous research has shown that item responses to these scales meet IRT modeling assumptions fairly well (Reise & Waller, 1990, 1993).

Real-Data Sample and Item Calibration

The "real-data" sample consisted of 2,000 persons drawn randomly from the Minnesota Twin Registry; this is the same sample used in Reise & Waller (1990). The sample consisted of 1,127 females [mean age = 40.78, standard deviation (SD) = 9.36] and 873 males (mean age = 43.09, SD = 10.60). IRT item parameters were calibrated using the 2,000 response vectors from the real-data sample with the MULTILOG computer program (Thissen, 1986). The 2PLM with $d = 1$ (Hambleton, Swaminathan, & Rogers, 1991, p. 15) was selected and all default MULTILOG conditions were used. For each item, MULTILOG provided estimates of the item discrimination (a) and item difficulty (b) parameters. These parameter estimates were treated as true population values in order to generate simulated response vectors in Studies 1 and 2.

The l_z Statistic

Given that an examinee has responded to a set of items with known parameters, θ can be estimated by finding the maximum of the likelihood function for the 2PLM. The resulting ML $\hat{\theta}$ represents the examinee's most likely position on the θ continuum given their pattern of responses. The likelihood of the response pattern conditional on θ (in log units) is

$$L|\theta = \sum_i L_i = \left[U_i \times \ln\left(P_i|\theta\right) + \left(1 - U_i\right) \times \ln\left(Q_i|\theta\right)\right]. \quad (1)$$

The expected value of $L|\theta$ is

$$\mathrm{E}\left(L|\theta\right) = \sum_i \left[P_i|\theta \times \ln\left(P_i|\theta\right) + Q_i|\theta \times \ln\left(Q_i|\theta\right)\right], \quad (2)$$

and the variance is

$$V(L|\theta) = \sum_i (P_i|\theta)(Q_i|\theta)\left[\ln(P_i|\theta/Q_i|\theta)\right]^2, \quad (3)$$

where $P_i|\theta$ is the probablity of a correct/keyed response, $Q_i|\theta = 1 - P_i|\theta$, and $U_i = 1$ for a correct/keyed response and 0 otherwise. Combining Equations 2 and 3, the Drasgow et al. (1985) standardized person-fit index, l_z, is,

$$l_z|\theta = L|\theta - E(L|\theta)/\left[V(L|\theta)\right]^{1/2}. \quad (4)$$

The conditional null distribution for l_z is the standard normal (Drasgow et al., 1985). Thus, l_z has an expected value of 0.0 and a variance of 1.0 conditional on θ. Large negative l_z values indicate nonfitting response patterns given θ; large positive l_z values indicate response patterns that are higher in likelihood than the model predicts; 0.0 is the expected value when an examinee's responses conform to the probabilistic IRT measurement model. Note that this normal distributional assumption for l_z rests on the central limit theorem and would only be expected to hold when the number of items is large. The present study, which used scales with a relatively small number of items, provides a test of the robustness of the l_z null distribution for relatively short tests.

Procedure

For each MPQ scale, two analyses were performed. In the first and second analyses, 10,000 response vectors were generated by treating the item parameters estimated from MULTILOG as true item parameters. The true θs were specified to be 2,000 constants at each of five θ levels: θ = −2.0, −1.0, 0.0, 1.0, and 2.0. These values were selected because (1) they represent a large range of θ, (2) it was important to examine the l_z distribution conditional on various θ levels to search for systematic differences, and (3) values of θ greater than 2.0 or lower than −2.0 would have resulted in many response vectors with all 1 or all 0 responses.

For each of the 2,000 simulated response vectors at each of the five θ levels per scale, l_z was computed twice. In the first analysis, l_z was computed using all the generating parameters (i.e., the true item and person parameters). In the second analysis, a ML $\hat{\theta}$ was computed for each response vector. These $\hat{\theta}$s, along with the true item parameters, then were used in the computation of l_z.

Results

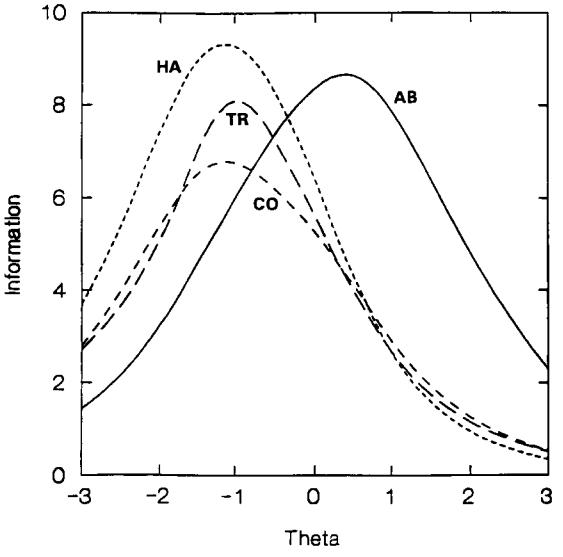
Descriptive statistics summarizing the item parameter estimates are displayed in Table 1. The average *b* was in the low θ range for the TR, HA, and CO scales. For AB, the average *b* was near the mean θ level (i.e., 0.0). The *a*s averaged approximately 1.1 within scales and they ranged from a low of .56 to a high of 2.67. Figure 1 shows the test information functions (TIFs; Hambleton et al., 1991, p. 94) for each scale. The TIFs show that all scales had peaked information functions, indicating a disproportionate amount of measurement precision within specific θ ranges. HA, TR, and CO provided the most information in the low θ range. The AB scale had peaked information in the middle of the θ continuum.

Table 1
Descriptive Statistics for Item Parameters

Scale and Parameter	Mean	SD	Maximum	Minimum
CO				
a	1.13	.38	2.00	.56
b	−.88	.82	.36	−2.41
HA				
a	1.18	.30	1.89	.62
b	−1.19	.53	−.23	−2.39
TR				
a	1.07	.43	2.67	.63
b	−1.10	.85	1.25	−3.19
AB				
a	1.07	.21	1.66	.75
b	.24	.89	2.09	−1.54

True item parameters and true θ. Table 2 provides summary statistics for the l_z null distributions conditional on θ when true item and person parameters were used to generate data and to compute l_z. It is evident that 0.0 was the average l_z value across scales, except when θ = 2.0, where the mean l_z was negatively biased for each scale except AB. This result may be due to the fact that the skew of l_z when

Figure 1
Test Information Functions for the MPQ Scales



$\theta = 2.0$ was especially large and negative for CO, HA, and TR, but not for AB; negative skews led to means that were shifted to the left.

A second notable finding was that the variances of the conditional l_z null distributions were approximately 1.0 in most conditions (see Table 2). The exceptions were when $\theta = 1.0$ for HA (variance = .81) and when $\theta = 2.0$ for CO, HA, and TR (variance = .66, .53, and .74, respectively). In the high θ range, the null distribution variances of l_z were truncated relative to a standard normal distribution for CO, HA, and TR.

Finally, the conditional l_z null distributions were always skewed negatively; for each scale and θ level, the maximum l_z score was constrained in comparison to the minimum l_z score. This result means that even in simulated data when all parameters are correct, the θ conditional l_z null distributions were not symmetric. Because in theory the index is standard normal given θ when the data fit the model, this was an important finding.

Inspection of the TIFs in Figure 1 helps clarify the findings. HA had the lowest information at $\theta = 2.0$, and the variance of the l_z distribution conditional on $\theta = 2.0$ was smallest for this scale. Figure 2a shows a plot of test information given θ ($I|\theta$) (from Figure 1) on the abscissa and the conditional variances of l_z (Table 2) on the ordinate. The relationship appears moderate but not perfect. As noted in Drasgow (1982), the l_z equations were meant to be good approximations to "standardizing" transformations only when $I|\theta$ is high. Perhaps under the present conditions, at high θ levels CO, HA, and TR did not have sufficient $I|\theta$.

True item parameters and estimated ML $\hat{\theta}$. Table 3 shows descriptive statistics for the conditional l_z null distributions based on true item parameters and ML $\hat{\theta}$ used in the computation of l_z. These results were similar to the results in Table 2 but differed in one major respect: when $\hat{\theta}$ was used in the computation of l_z, the conditional l_z null distributions were constrained in variability (i.e., < 1.0).

The reduction in variance appeared to depend on the $I|\theta$ within each scale. In Figure 2b, the variances of the l_z distributions (Table 3) are plotted as a function of the $I|\theta$ (from Figure 1). As in Figure 2a, the higher $I|\theta$, the closer the conditional l_z variance was to 1.0. This finding reinforces the notion that the variances of the l_z null distributions depend on some critical amount of $I|\theta$; if there is little information, there is little variance in the l_z null distributions. Note, however, that at some of the θ levels many of the simulated response vectors were all 0s or 1s (e.g., at $\theta = 2.0$ there were 297 vectors with all 1s for TR, 321 for CO, and 745 for HA). When this occurred l_z was set to 0.0, which constrained the variance values reported in Table 3.

The relationship between $I|\theta$ and the conditional l_z variance shown in Figure 2b was not perfect, however. For example, Table 3 shows that when $\theta = -1.0$, the variance for CO (.96) was much closer to 1.0 than for AB (.43). However, at this low θ level CO had only slightly more $I|\theta$ than AB. This result indicates that it was not the amount of $I|\theta$ that solely accounted for the conditional l_z variances. What appeared to be critically important in determining the variances of the l_z null distributions was the number of items within each scale that had difficulties near a given θ level.

Equation 3 reveals that the potential range of l_z values can be influenced depending on whether a scale contains items for which the examinee has a

Table 2
Descriptive Statistics of Conditional Null Distributions of l_z When True θ Was Used to Compute l_z

θ and Scale	Mean	Variance	Skewness	Kurtosis	Maximum	Minimum
θ = −2.0						
CO	−.01	.99	−.48	.07	2.17	−3.91
HA	−.01	.93	−.42	.25	2.38	−4.18
TR	0.00	1.00	−.55	.35	2.20	−5.10
AB	−.06	.97	−.73	.65	1.72	−4.74
θ = −1.0						
CO	−.02	1.01	−.42	.02	2.35	−3.59
HA	.02	1.02	−.37	.10	2.39	−4.06
TR	0.00	1.03	−.48	−.03	2.29	−3.73
AB	−.02	1.00	−.62	.66	2.36	−5.13
θ = 0.0						
CO	0.00	1.02	−.89	1.22	1.84	−5.59
HA	.01	.95	−.43	−.01	2.37	−3.40
TR	−.04	1.00	−.42	.20	2.41	−4.48
AB	−.02	.97	−.35	.13	2.68	−3.62
θ = 1.0						
CO	−.03	.98	−.88	1.32	1.68	−5.14
HA	−.09	.81	−.94	1.26	1.27	−5.05
TR	−.04	.98	−.63	.48	1.78	−4.41
AB	.04	.95	−.44	.09	2.35	−3.58
θ = 2.0						
CO	−.19	.66	−1.42	3.32	.93	−5.74
HA	−.34	.53	−2.27	6.11	.37	−5.25
TR	−.19	.74	−1.25	2.12	1.02	−4.47
AB	0.00	.99	−.54	.44	2.12	−4.33

$P|\theta$ near .5. As $P|\theta$ goes from 0 or 1 toward .5, values in Equation 3 decrease, which makes any value in the numerator of l_z result in a larger l_z. Returning to the above example, although AB contained

Figure 2
$I|\theta$ Versus the l_z Variance$|\theta$ Using True θ and Estimated θ

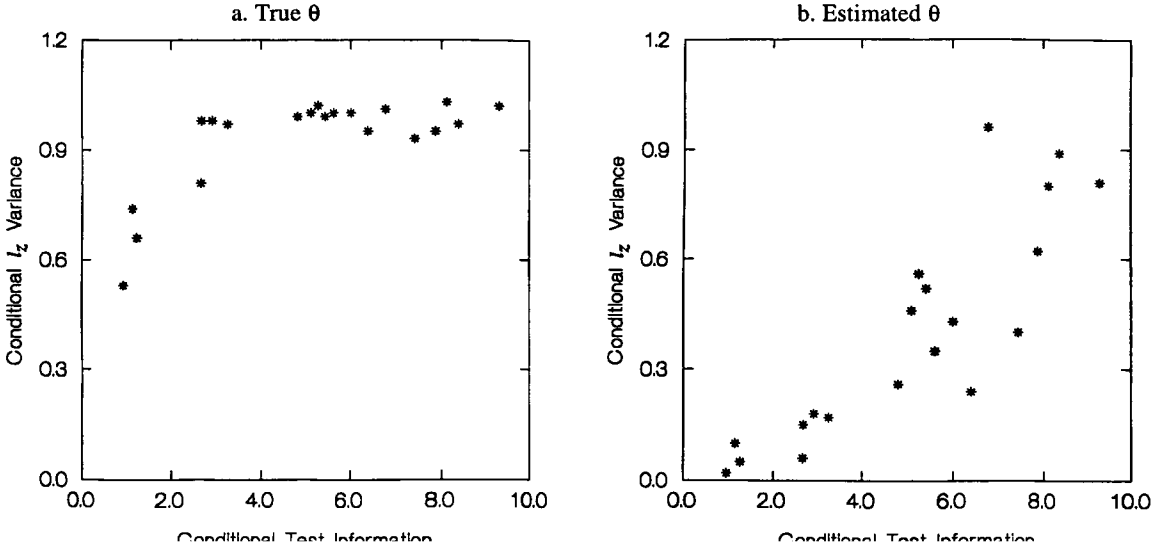


Table 3
Descriptive Statistics of Conditional l_z Distributions When θ was Estimated With ML

θ and Scale	Mean	Variance	Skewness	Kurtosis	Maximum	Minimum
θ = −2.0						
CO	.05	.52	−.30	.12	2.40	−3.04
HA	.04	.40	−.50	1.28	2.17	−3.78
TR	.05	.46	−.59	.62	1.84	−3.10
AB	.02	.17	−.42	.40	1.18	−1.67
θ = −1.0						
CO	.11	.96	−.47	.30	2.40	−4.26
HA	.06	.81	−.35	0.00	2.83	−3.22
TR	.04	.80	−.37	.12	2.65	−3.54
AB	.01	.43	−.39	.49	1.75	−3.44
θ = 0.0						
CO	.07	.56	−.47	.91	2.12	−4.07
HA	.01	.24	−.57	3.58	2.25	−3.92
TR	.03	.35	−.43	.26	1.78	−2.55
AB	.08	.89	−.33	.18	2.64	−4.13
θ = 1.0						
CO	.02	.18	−.60	2.07	1.53	−2.54
HA	0.00	.06	−.20	.88	.91	−1.13
TR	.01	.15	−.49	.24	1.17	−1.72
AB	.06	.62	−.30	.20	2.20	−3.40
θ = 2.0						
CO	.01	.05	−.82	1.89	.76	−1.12
HA	0.00	.02	−.24	.99	.66	−.66
TR	.02	.10	−.16	−.20	.85	−1.28
AB	.02	.26	−.28	.29	1.66	−2.16

no items that discriminated best at θ = −1.0, it had almost as much test information at θ = −1.0 as CO due to its larger number of items. However, the CO scale contained many items with *b*s of approximately −1.0 (see Table 1). Hence, for CO, items can be administered for which examinees with θ = −2.0 would have $P|\theta$ near .5. This property would allow l_z to approach its asymptotic variance of 1.0. Thus, information is important, but having items with θ = *b* is of equal influence in determining the conditional l_z null distributions.

Study 2: Detection Power and Scoring Method

Study 2 was concerned with the power of l_z in detecting non-model-based responding and how detection power is affected by scoring method. There are several fundamental problems in addressing the detection "power" issue. For example, the power of the statistic to detect nonfit potentially interacts with a number of factors, such as (1) the degree of *a*, (2) the distribution of *b*s, (3) the number of items, (4) the scoring method, and (5) the type of nonfit. It is difficult to address all these variables in one study; hence, some factors were held constant and others were manipulated.

Preliminary Issues

The primary question that arises in any attempt to detect nonfitting response vectors is the type of nonfit the statistic will detect. Person-fit researchers (e.g., Drasgow, 1982) have typically examined the power of various fit statistics to detect one or two types of well-defined lack of fit (e.g., a high θ examinee mismarks 10% of the items or a low θ examinee cheats on 10% of the items).

In the personality assessment context, there are a variety of types of unusual responding, such as lack of attention to particular items, randomly responding to an inapplicable item, and deliberate faking. It is difficult to specify with precision how these factors may affect the probability of endorse-

ment across a range of individuals. In other words, there is a multitude of potential non-model-based faulty person-item interactions that may or may not combine to produce a statistically non-model-fitting response pattern, but certainly do contribute to a test score being a less precise indicator of θ level. Therefore, the power of a person-fit statistic to detect defined forms of non-model-fit is clearly an endless research question.

A more empirically tractable issue is the power of l_z to detect non-model-based responses without regard to their specific etiologies. The specific types of misfit mentioned above have a common result—they all produce a response that is not derived (i.e., predicted) from the unidimensional IRT model. Hence, in this study, lack of model fit was defined as any response that was not generated from a specified unidimensional 2PLM. Non-model-based responses were generated here by first specifying a set of item parameters, manipulating the *a* parameters to some value lower than that specified by the model, and then generating item responses. The resulting response vectors were then, on average, "lower" in likelihood than expected from the original parameters, and then were tested for nonfit with respect to the model specified by the original item parameters (Reise & Due, 1991).

In Study 2, the *a* parameters taken from the four MPQ scales were systematically set to 0.0 in order to generate simulated nonfitting response vectors. The effects of setting $a = 0.0$ and then generating a response were: (1) the response becomes random (i.e., there is a .5 probability of endorsement or nonendorsement), and (2) the simulated response provides no psychometric information toward estimating θ because $a = 0.0$. Within a scale, any number of *a*s can be set to 0.0 to simulate different frequencies of nonfitting responses. Thus, the issue of the power of l_z to detect nonfit as a function of the number of nonfitting responses can be addressed. Of course, l_z was always calculated assuming the actual MPQ item parameters, as would be done in real testing.

A second salient issue was whether the detection of nonfit should be studied at particular θ levels (i.e., conditionally) or be examined within a specified distribution of θ levels. Gafni (1988) explicitly took the former approach and most others have taken the latter. As in Study 1, detection power was investigated at a few distinct θ levels. This approach allowed for empirical comparisons of how detection power changed as a function of θ.

Method

Estimation of θ

ML estimation. A ML estimate of θ is determined by summing the log likelihood of each observed item response (see Equation 1) conditional on θ. Then, the maximum (i.e., the mode) of the likelihood function must be determined (Lord, 1980). In this study, an iterative Newton-Raphson procedure was used to find the mode of each examinee's likelihood function. Response vectors containing all 1s (all endorsed) or all 0s (all not endorsed) were assigned θ estimates of 3.0 and −3.0, respectively.

The ML θ has several positive asymptotic features (Hambleton et al., 1991). First, it is not biased (i.e., the expected value of $\hat{\theta}$ always equals θ). Furthermore, it is efficient and normally distributed. However, the ML estimator has some problems. No ML estimate can be obtained from response vectors with all 0s or 1s. Furthermore, these properties depend on the assumption that the responses fit the model (Mislevy & Bock, 1982). Under conditions of nonmodel-fit and finite test lengths, there is no guarantee that these statistical properties will be maintained.

Expected a posteriori estimation. In contrast to the ML estimator, expected a posteriori (EAP) estimation (Bock & Mislevy, 1982) is noniterative and provides a finite $\hat{\theta}$ for all response patterns. The EAP is a Bayesian estimator derived from finding the mean of the posterior distribution (Bock & Mislevy, 1982, p. 433). 61 quadrature nodes were used ranging from $\theta = -3.0$ to 3.0 in .10 increments.

Bock & Mislevy (1982) stated "the EAP estimator has minimum mean square error over the population of ability and, in terms of average accuracy, cannot be improved upon" (p. 439). This property only applies when the prior is correct, however (Wainer & Thissen, 1987). The EAP estimator is biased when there is a finite number of items; that is, $\hat{\theta}$ is regressed toward the mean unless the num-

ber of items is large (Wainer & Thissen, 1987) and how large is "large" is unknown.

Biweight estimation. The biweight (BIW) (Mislevy & Bock, 1982) estimate of θ is a robust estimator (Wainer & Thissen, 1987). This means that it is supposed to ignore or "downweight" unusual item responses when computing $\hat{\theta}$; thus, it can potentially provide a better estimate of an examinee's position on the θ continuum when there are unusual or nonfitting responses. Computation of the BIW proceeded in two stages.

In the first stage, the ML $\hat{\theta}$ was determined using the Newton-Raphson procedure. Then, the ML estimate was used as the starting value for the second stage of the Newton-Raphson iterations in which the actual BIW was estimated. To compute BIW, a weight must be determined for each item response. Once determined, the weight was then multiplied into the first derivative of the likelihood function during the Newton-Raphson iterations. The number of iterations to compute the BIW was fixed at 2 to circumvent the "run-away" parameter problem discussed in Mislevy & Bock (1982).

For each item response, the weight was set to $(1 - K^2)^2$ if the absolute value of K was less than 1.0, and 0 otherwise. The K values were computed using

$$K = a\left(b - \hat{\theta}\right)\Big/4.0 \qquad (5)$$

(Mislevy & Bock, 1982). The weight is largest when $b = \hat{\theta}$ and the weight can never be greater than 1.0. As b moves away from $\hat{\theta}$, the item response is downweighted. If the discrepancy between b and $\hat{\theta}$, multiplied by a, is greater than 4.0, then K has an absolute value greater than 1.0. When $K > 1.0$, the weight is set to 0.0 (i.e., the first derivative is multiplied by 0.0); thus, the item response does not contribute to estimation of θ.

Mislevy & Bock (1982) noted several advantages to BIW. A major advantage is that the item responses are used in proportion to their potential value. This means that items close to an examinee's $\hat{\theta}$ are weighted highly and make a large contribution toward estimating θ. Items that are far away in difficulty from the examinee's $\hat{\theta}$ make little or no contribution toward estimating θ. The rationale behind the development of BIW was that "maximum likelihood estimates ... are overly sensitive to measurement disturbances that are common in educational testing..." (Mislevy & Bock, 1982, p. 725).

Procedure

All monte carlo simulations were performed using computer programs written by the author. The simulations were conducted in cycles that repeated $n + 1$ times per scale, where n is the number of scale items. In the first cycle, 1,000 response vectors were simulated for each of three θ levels: −1.5, 0.0, and 1.5. Then, for each response vector, l_z was computed four ways: (1) using the true (generating) θ value (called TRUE scoring), (2) using the ML $\hat{\theta}$, (3) using the EAP $\hat{\theta}$, and (4) using the BIW $\hat{\theta}$. The TRUE method of computing l_z (i.e., using all true generating person and item parameters) was a control condition because detection rates should be maximum when all the parameters used in the computation of the statistic are error-free. This unrealistic condition served as a baseline for evaluating the three $\hat{\theta}$ conditions.

After the first cycle, in which the number of simulated nonfitting item responses was 0, the entire simulation cycle was repeated but with one item designated to generate non-model-based responses. This was accomplished by setting the generating a parameter for the designated nonfitting item to 0.0. l_z was computed, as before, using the true generating item parameters (i.e., no items with 0.0 as) and the four scoring methods. This cycle of generating nonfitting response vectors at three θ levels and computing l_z with the different scoring methods then was repeated with two items with $a = 0.0$, with three items with $a = 0.0$, and so on until all the items in the scale had generating as of 0.0. The order of the items designated to have 0.0 generating as was the same order that the items appeared in the MPQ test booklet.

Within each cycle of computing l_z four ways for the 1,000 response vectors at each of the three θ levels, each observed l_z value was evaluated for statistical significance by comparing it with three Z-score critical values from the standard normal distribution (i.e., the null distribution). The three

one-tailed critical values or false positive rates used were $\alpha = .01$, .05, and .10, which correspond to Z values of −2.32, −1.65, and −1.28, respectively.

If l_z was statistically significant at a given α level, it was counted as a "hit," otherwise it was considered a "miss." The hit rates were then defined as the number of hits divided by 1,000 at each of the three α levels. The final data then consisted of hit rates for l_z for four personality scales, three θ levels, four scoring methods, and $n + 1$ levels of nonfit, all at three significance levels.

Results

Tables 4–6 show hit rates for all scales and scoring conditions for $\theta = -1.5$, 0.0, and 1.5, respectively, at the $\alpha = .05$ significance level (results for all conditions are available from the author).

Hit Rates Using True θ

This analysis eliminated θ estimation as a source of error, and thus allowed for the identification of between-scale differences in detection

Table 4
Hit Rates at $\theta = -1.5$ for the Four Scales as a Function of the Number of Random Responses (k) and Scoring Method for $\alpha = .05$

	CO				HA				TR				AB			
k	TRUE	ML	EAP	BIW	TRUE	ML	EAP	BIW	TRUE	ML	EAP	BIW	TRUE	ML	EAP	BIW
0	.06	.03	.03	.04	.06	.03	.04	.04	.05	.03	.03	.03	.07	0.00	0.00	.03
1	.06	.03	.04	.05	.09	.04	.03	.05	.09	.05	.05	.06	.08	0.00	0.00	.04
2	.09	.05	.05	.06	.09	.04	.05	.04	.10	.06	.05	.05	.09	0.00	0.00	.05
3	.07	.05	.04	.05	.09	.05	.05	.06	.10	.04	.05	.06	.12	0.00	0.00	.05
4	.08	.05	.05	.06	.10	.05	.04	.05	.11	.06	.05	.05	.14	.01	0.00	.03
5	.09	.06	.06	.06	.08	.05	.06	.06	.19	.12	.10	.15	.30	.03	.02	.12
6	.14	.08	.10	.07	.09	.05	.06	.06	.28	.15	.15	.18	.30	.03	.03	.10
7	.13	.07	.09	.09	.14	.07	.07	.08	.28	.16	.17	.16	.35	.04	.02	.14
8	.16	.08	.08	.10	.14	.07	.08	.08	.35	.19	.19	.18	.44	.06	.05	.17
9	.19	.11	.12	.13	.13	.06	.08	.09	.38	.22	.23	.26	.57	.15	.14	.31
10	.23	.13	.12	.15	.16	.06	.08	.08	.44	.25	.26	.30	.69	.25	.20	.41
11	.21	.13	.13	.13	.17	.08	.07	.09	.41	.28	.28	.30	.77	.28	.23	.41
12	.27	.21	.18	.21	.19	.08	.09	.09	.42	.29	.25	.29	.81	.35	.30	.49
13	.48	.37	.38	.42	.21	.09	.09	.08	.52	.44	.39	.43	.89	.55	.49	.63
14	.54	.36	.36	.43	.21	.12	.09	.10	.53	.41	.41	.44	.91	.51	.49	.66
15	.63	.46	.45	.48	.39	.18	.22	.22	.57	.42	.42	.44	.91	.50	.46	.60
16	.69	.55	.54	.58	.43	.25	.26	.26	.61	.45	.43	.48	.94	.56	.55	.65
17	.71	.52	.54	.53	.46	.26	.26	.28	.63	.43	.46	.46	.95	.57	.50	.60
18	.78	.61	.56	.60	.51	.25	.28	.29	.64	.42	.48	.48	.96	.56	.50	.63
19	.77	.57	.61	.60	.50	.29	.30	.31	.64	.48	.48	.49	.96	.53	.50	.60
20	.80	.62	.60	.62	.58	.35	.32	.33	.63	.47	.47	.47	.97	.52	.51	.62
21	.80	.63	.62	.63	.57	.33	.34	.34	.63	.46	.45	.47	.97	.55	.56	.63
22	.80	.63	.64	.63	.57	.36	.35	.34	.67	.47	.48	.48	.97	.55	.56	.64
23	.82	.68	.68	.71	.61	.33	.33	.31	.72	.49	.51	.52	.98	.54	.52	.59
24	.83	.68	.69	.70	.57	.34	.38	.30	.77	.53	.52	.55	.98	.60	.59	.67
25					.67	.45	.47	.43	.78	.57	.59	.56	.99	.68	.66	.70
25					.67	.46	.50	.47	.79	.52	.57	.58	.99	.67	.69	.72
27					.66	.46	.50	.46	.80	.59	.61	.58	1.00	.73	.74	.79
28					.67	.48	.47	.45					1.00	.75	.71	.79
29													1.00	.81	.78	.80
30													1.00	.77	.77	.80
31													1.00	.83	.81	.87
32													1.00	.80	.80	.86
33													1.00	.82	.84	.88
34													1.00	.80	.77	.84

Table 5
Hit Rates at θ = 0.0 for the Four Scales as a Function of the Number of Random Responses (k) and Scoring Method for α = .05

	CO				HA				TR				AB			
k	TRUE	ML	EAP	BIW	TRUE	ML	EAP	BIW	TRUE	ML	EAP	BIW	TRUE	ML	EAP	BIW
0	.08	.02	.03	.04	.05	0.00	0.00	.01	.07	.01	.02	.03	.06	.05	.05	.06
1	.09	.03	.04	.05	.07	0.00	0.00	.01	.07	.01	.01	.02	.06	.04	.06	.05
2	.08	.03	.03	.05	.07	0.00	0.00	.01	.10	.02	.02	.04	.06	.04	.04	.05
3	.15	.05	.07	.09	.10	0.00	0.00	.01	.10	.02	.02	.03	.07	.04	.05	.06
4	.16	.06	.06	.10	.12	0.00	0.01	.02	.12	.01	.02	.02	.05	.03	.03	.06
5	.36	.13	.14	.19	.19	.01	.01	.02	.11	.03	.04	.04	.13	.11	.11	.11
6	.37	.13	.13	.20	.22	.02	.02	.03	.15	.03	.04	.05	.15	.13	.11	.13
7	.39	.14	.15	.21	.25	.02	.02	.05	.15	.03	.03	.05	.15	.11	.11	.14
8	.52	.21	.22	.31	.31	.03	.04	.05	.16	.02	.03	.05	.18	.15	.14	.14
9	.56	.20	.24	.31	.36	.02	.04	.05	.17	.03	.04	.05	.24	.20	.19	.21
10	.54	.21	.23	.27	.38	.01	.04	.07	.30	.08	.09	.11	.28	.20	.21	.23
11	.56	.21	.25	.30	.46	.02	.02	.07	.39	.11	.13	.17	.32	.24	.23	.23
12	.73	.44	.45	.53	.47	.03	.02	.08	.41	.12	.13	.18	.32	.23	.21	.25
13	.73	.45	.49	.54	.61	.06	.08	.15	.59	.33	.34	.36	.47	.36	.39	.35
14	.81	.49	.50	.62	.66	.09	.10	.22	.62	.32	.34	.40	.47	.35	.35	.36
15	.80	.50	.49	.58	.70	.07	.08	.21	.74	.37	.40	.47	.51	.42	.43	.44
16	.81	.49	.52	.60	.71	.07	.08	.23	.76	.38	.40	.44	.53	.42	.42	.45
17	.83	.51	.52	.62	.74	.06	.08	.24	.78	.37	.40	.52	.52	.46	.45	.47
18	.83	.47	.52	.59	.75	.07	.08	.27	.80	.43	.46	.54	.54	.46	.43	.47
19	.88	.58	.57	.68	.76	.05	.08	.28	.81	.43	.43	.56	.57	.50	.50	.49
20	.88	.55	.56	.68	.78	.06	.06	.28	.84	.45	.45	.57	.62	.54	.56	.58
21	.90	.60	.58	.74	.83	.08	.10	.37	.86	.48	.49	.61	.63	.57	.54	.56
22	.90	.59	.61	.74	.84	.09	.12	.43	.88	.51	.53	.65	.64	.57	.55	.57
23	.94	.70	.70	.80	.90	.11	.13	.47	.89	.48	.50	.66	.65	.58	.56	.58
24	.95	.71	.71	.82	.90	.13	.15	.52	.90	.46	.48	.66	.68	.60	.60	.62
25					.96	.35	.35	.67	.92	.47	.54	.68	.71	.63	.63	.65
26					.98	.39	.41	.75	.93	.50	.53	.69	.74	.67	.64	.67
27					.97	.45	.45	.76	.96	.59	.60	.78	.74	.66	.66	.66
28					.99	.45	.45	.81					.74	.66	.66	.69
29													.79	.70	.69	.70
30													.77	.71	.70	.70
31													.80	.73	.70	.72
32													.82	.74	.71	.72
33													.84	.75	.76	.75
34													.87	.80	.80	.80

hit rates under optimal circumstances. An interesting result was evident for these TRUE results: Hit rates were optimal when the θ used to generate the response vectors was distant from the average b within a given scale. For example, in Table 6 hit rates in the AB scale were clearly lower than those for HA; when 6 nonfitting responses were generated, l_z identified 70% of the vectors as nonfitting in HA, but only 24% were so identified in AB. However, AB had the most $I|\theta$ at θ = 1.5 and HA had the least.

This result appeared to be a function of an interaction between the within-scale item parameters and the nonfit generation technique used in Study 2. Generating random responses (i.e., nonfitting responses) makes a vector extremely unlikely at θ = 1.5 in HA because the HA items had bs in the low θ range. For examinees at θ = 1.5, the HA item parameters specify probabilities of endorsement ($P|\theta$) of approximately .9. The simulated nonfitting responses had $P|\theta = .5$. The difference between .9 and .5 in the probability of endorsement

Table 6
Hit Rates at θ = 1.5 for the Four Scales as a Function of the Number of Random Responses (*k*) and Scoring Method for α = .05

	CO				HA				TR				AB			
k	TRUE	ML	EAP	BIW	TRUE	ML	EAP	BIW	TRUE	ML	EAP	BIW	TRUE	ML	EAP	BIW
0	.05	0.00	0.00	.04	.04	0.00	0.00	.04	.07	0.00	0.00	.04	.06	.01	0.00	.03
1	.11	0.00	0.00	.05	.08	0.00	0.00	.04	.14	0.00	0.00	.09	.09	.02	.01	.03
2	.12	0.00	0.00	.04	.14	0.00	0.00	.06	.26	0.00	0.00	.16	.08	.02	.01	.03
3	.32	.01	.01	.24	.28	0.00	0.00	.13	.28	0.00	0.00	.16	.11	.02	.02	.04
4	.44	.02	.02	.36	.42	0.00	0.00	.16	.38	0.00	0.00	.18	.15	.03	.02	.05
5	.71	.18	.16	.64	.55	0.00	0.00	.29	.36	0.00	0.00	.18	.14	.04	.02	.05
6	.78	.15	.15	.66	.70	0.00	0.00	.44	.52	0.00	0.00	.30	.24	.07	.06	.12
7	.79	.15	.15	.65	.82	0.00	0.00	.55	.56	0.00	0.00	.31	.25	.09	.06	.11
8	.86	.34	.33	.77	.83	0.00	0.00	.60	.69	0.00	0.00	.45	.28	.07	.06	.11
9	.90	.31	.32	.77	.89	0.00	0.00	.69	.75	0.00	0.00	.48	.27	.09	.06	.11
10	.92	.35	.36	.79	.92	0.00	0.00	.77	.84	.01	.01	.62	.29	.09	.05	.10
11	.93	.36	.37	.84	.95	0.00	0.00	.82	.91	.02	.03	.71	.34	.09	.06	.12
12	.96	.60	.63	.92	.96	0.00	.01	.84	.92	.03	.03	.77	.34	.07	.07	.13
13	.97	.56	.56	.92	.98	.01	.03	.92	.96	.23	.23	.85	.35	.11	.09	.14
14	.99	.63	.63	.95	1.00	.05	.05	.93	.98	.23	.24	.89	.41	.12	.09	.15
15	.99	.55	.54	.97	.99	.03	.04	.95	.98	.30	.31	.94	.58	.24	.20	.30
16	.99	.50	.51	.97	1.00	.03	.03	.96	.99	.31	.32	.94	.62	.29	.23	.35
17	.99	.52	.56	.98	1.00	.02	.03	.98	1.00	.32	.32	.97	.72	.36	.32	.43
18	1.00	.48	.51	.98	1.00	.04	.03	.98	.99	.39	.43	.97	.76	.40	.33	.47
19	1.00	.60	.60	.99	1.00	.02	.03	.99	1.00	.38	.40	.97	.83	.50	.48	.59
20	1.00	.56	.56	.99	1.00	.01	.03	.99	1.00	.43	.46	.98	.88	.57	.56	.68
21	1.00	.62	.61	.99	1.00	.03	.04	.99	1.00	.51	.52	.99	.90	.59	.59	.69
22	1.00	.61	.62	1.00	1.00	.04	.06	1.00	1.00	.54	.54	.99	.92	.65	.62	.74
23	1.00	.75	.73	1.00	1.00	.06	.07	1.00	1.00	.47	.51	1.00	.94	.68	.70	.80
24	1.00	.71	.69	1.00	1.00	.08	.08	1.00	1.00	.43	.43	.99	.95	.70	.68	.81
25					1.00	.33	.33	1.00	1.00	.47	.46	1.00	.94	.67	.68	.80
26					1.00	.38	.42	1.00	1.00	.49	.50	1.00	.95	.70	.69	.80
27					1.00	.44	.48	1.00	1.00	.59	.59	1.00	.96	.70	.68	.81
28					1.00	.45	.47	1.00					.97	.69	.70	.82
29													.97	.69	.69	.83
30													.97	.70	.69	.83
31													.98	.70	.68	.83
32													.99	.72	.72	.90
33													.98	.70	.70	.89
34													.99	.79	.82	.94

is large, which leads to high rates of nonfit detection in HA under TRUE scoring.

However, $P|\theta$ for AB conditional on $\theta = 1.5$ hovered around .6 across items. Thus, the simulated nonfitting responses, which had $P|\theta = .5$ of endorsement, could not be readily distinguished because the discrepancy between true and nonfitting response probabilities was not great. The finding that the detection of nonfit was best in the θ range away from the average *b* leads to a paradox.

To measure a person with precision, items must be administered that match the examinee's θ level (Weiss, 1982). To identify lack of fit well, items must be administered for which the examinee has very low or very high probability of endorsing. In this way, nonfitting (i.e., random) responses can be readily identified as being non-model-generated. Concisely, when $\theta = b$ a random vector of responses is the expected response pattern. Hence, because of the type of nonfit imposed on the data in this study, the power to detect nonfit was maximized at θ ranges far from the average of the *b*s

within each scale.

The interaction between the data generation and the *b*s makes it difficult to comparatively evaluate the four MPQ scales in regard to detection of nonfit. Under TRUE scoring, the most that can be said is that for $\theta = -1.5$ (Table 4), detection of nonfit was best for AB. When $\theta = 0.0$ (Table 5), hit rates were approximately the same across scales. For $\theta = 1.5$ (Table 6), detection of nonfit was superior in CO, HA, and TR.

Hit Rates for ML Scoring

Figures 3a–3d show hit rate plots at the three levels of θ for CO, HA, TR, and AB, respectively. Only the results for ML scoring are shown because they adequately illustrated the effects of θ level on detection power for a frequently used scoring method, and the results for EAP estimation were nearly identical to the ML results (effects of BIW scoring are discussed below).

From Figure 3a, which illustrates detection in the CO scale, it is apparent that detection was best when $\theta = 1.5$. However, once there were approximately 16 nonfitting responses, detection power was approximately the same across the three θ levels. In the CO scale, examinees must respond to at least 12 items randomly at $\theta = 1.5$ but about 15 items at $\theta = 0.0$ and $\theta = 1.5$ in order for l_z to have a .5 probability of detecting a random response pattern. l_z was able to identify approximately 70% of completely random response vectors as significantly nonfitting when all the responses did not fit.

Figures 3b and 3c show a different effect. In HA and TR, detection was best at $\theta = -1.5$. The HA and TR scales had average *b*s in the low θ range (see Table 1) and this was where nonfit was detected best under ML scoring. This finding is contrary to the results that indicated detection power was best in the θ range away from the average *b* when TRUE scoring was used. Thus, for HA and TR, the data indicate that random responding was difficult to detect in these scales. For HA, less than 50% of the cases could be identified as random when all the responses were generated to be nonfitting. For TR, 60% detection was the maximum for the completely nonfitting response conditions.

Figure 3d shows the hit rates for the AB scale. In this scale, it is not clear at which θ level nonfit detection was best because the curves cross several times. Across the θ levels however, detection rates appeared fairly good in this scale. Approximately 20 random responses were needed to have a .5 probability of identifying the examinee as nonfitting. Over 75% of examinees who responded randomly to all the items were identified with l_z as significantly nonfitting.

Note that in Figures 3a–3d when the number of random responses equaled the number of scale items (*n*) in each scale, the hit rates were approximately the same within scales at all true θ levels. This was expected because when all the items were simulated with $a = 0.0$, the true θ no longer affected the probability of any response within a scale. This would lead to the generated matrices being randomly equivalent across θ level conditions whenever the number of random responses equaled *n*. When all the item responses were simulated to be random, it did not make any difference what the θ level was when computing hit rates within a scoring method.

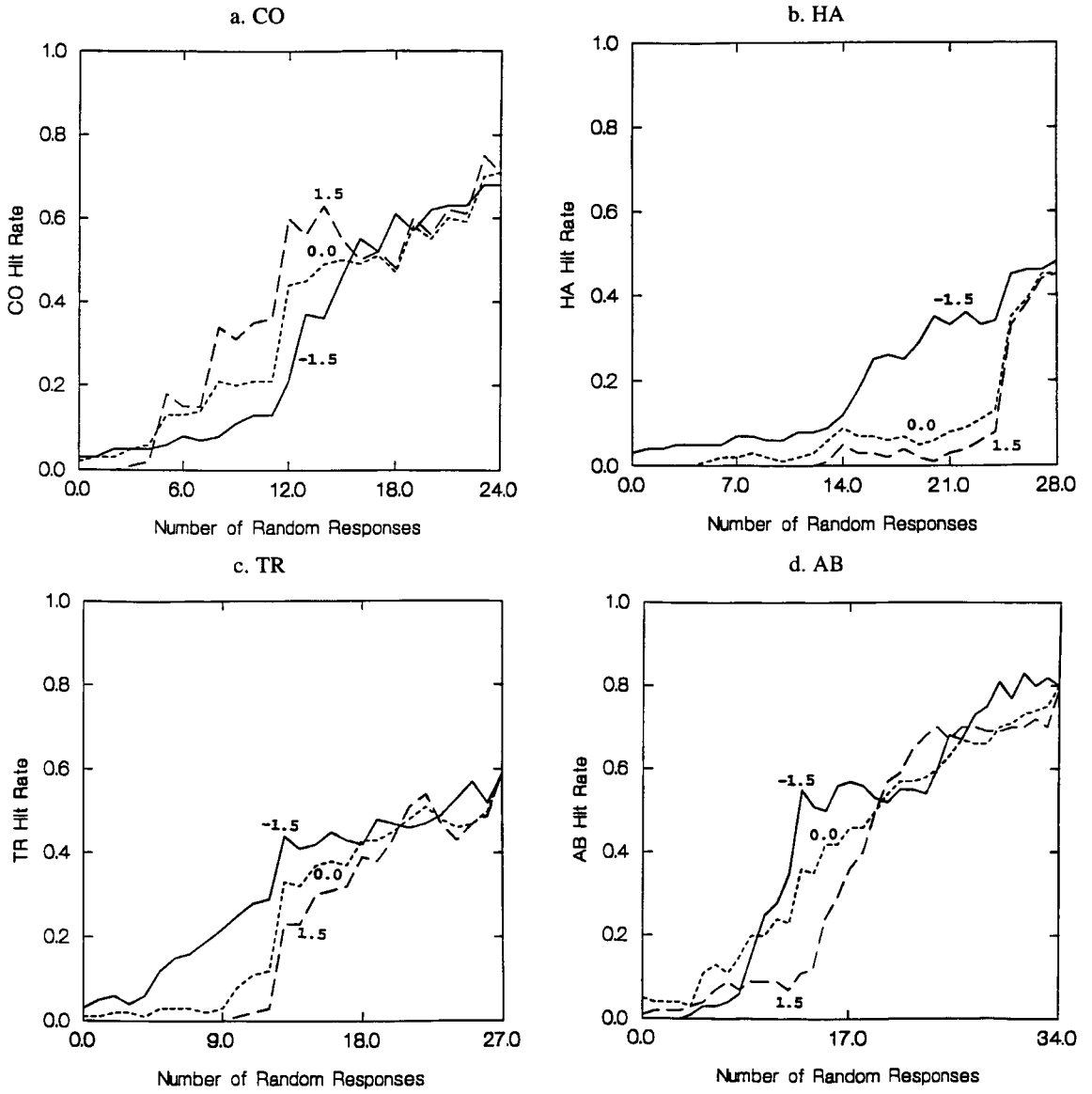
Hit Rates Across Scoring Methods

Figures 4a–4d show the hit rates as a function of scoring method when $\theta = 0.0$ for each of the MPQ scales, respectively (results for $\theta = -1.5$ and 1.5 were similar). As is evident in Figures 4a–4d, hit rates were always diminished by the estimation of θ. Clearly, the proportions of nonfitting response patterns identified as nonfitting were never greater in an estimated θ condition than in the corresponding TRUE scoring condition. This result supports the contention made earlier: l_z is optimal in detecting nonfit only when true θ is used.

A second major result was that BIW scoring resulted in the best hit rates of the three estimation methods in the CO, HA, and TR scales. For AB (Figure 4d), the scoring method did not appear to have much effect on detection rates. These results are interesting because they indicate that a robust scoring method can be coupled with l_z to result in higher nonfit detection rates.

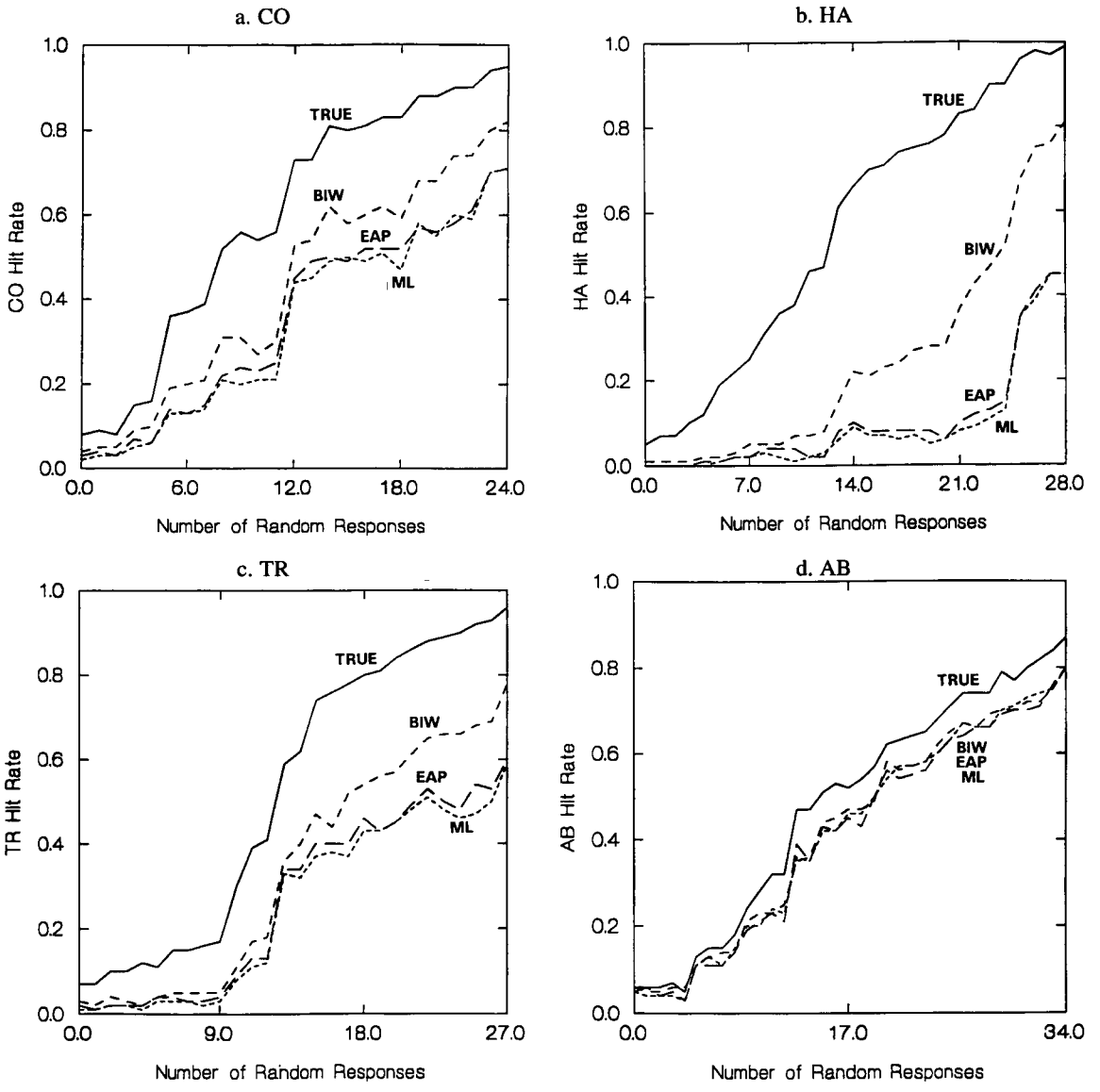
For example, for TR under the TRUE scoring condition, hit rates were highest when $\theta = 1.5$ (Table 6)

Figure 3
Number of Nonfitting Responses Versus the Hit Rates for Each Scale for Three θ Levels When ML Scoring was Used



and lowest when θ = −1.5 (Table 4). Likewise, with BIW scoring hit rates were highest when θ = 1.5 and lower when θ = 0.0 or −1.5. That is, detection was maximum away from the average *b* within scales. By contrast, for ML scoring detection was greatest at the θ level near the average *b* for the HA and TR scales. Perhaps this finding was a consequence of ML not being as good an estimator as BIW under conditions of nonfit. If this is true, then it could be concluded that nonfit detection hit rates are maximized when an examinee receives items away from their true θ level, but only if a good estimate of θ can be obtained. It appears from these results that ML was not a good θ level indicator in the HA and

Figure 4
Number of Nonfitting Responses Versus the Hit Rates for Each Scale for TRUE and for three Methods of θ Estimation



TR scales, and the BIW robust estimator should be used when nonfit is suspected because it appears to be less sensitive to random responding.

A third finding was that it did not appear to make any difference, as far as hit rates are concerned, whether ML or EAP scoring was used. Figures 4a–4d show that the EAP and ML curves nearly overlapped within scales for CO, HA, and TR. Again, however, in the AB scale no scoring method was any better than any other (except TRUE, of course). This result, when combined with the ML results for the AB scale, leaves the impression that when the number of items is large (i.e., 34) and information is spread over the θ range, it does not seem to matter which

scoring method is used, or what the examinee's θ level is. In the CO, HA, and TR scales, however, scoring method and θ level appear to be important variables.

Discussion

Standardization of l_z

Previous l_z research conducted in the ability domain (e.g., Drasgow et al., 1985, 1987) supported the contention that l_z is distributed approximately as standard normal when examinees respond according to a specified IRT model. The results obtained here clearly indicated that l_z was not well standardized across all θ levels, especially when θ was estimated. Because the l_z null distribution variances were often less than 1.0 and changed as a function of θ level within scales, the results lead to the conclusion that the standardization formulas presented in Equations 2–4 were incorrect when applied in the present context.

The plots of $I|\theta$ versus θ conditional l_z variance across scales showed that some critical level of $I|\theta$ and having item difficulties near a particular θ level are both important factors influencing whether the l_z standardizing equations produce the desired distributional results. Hence, an interesting finding came from empirically determining the l_z null distributions. That is, it is not $I|\theta$ that solely determines whether the l_z conditional null distribution variances will reach their asymptotic value of 1. The number of items with difficulties near a particular θ level is also an important feature.

For applied situations, finding that l_z was not distributed standard normal across θ levels might be tolerable. Finding that the null distributions were not consistent across θ levels is problematic. If the null distributions are consistent, but not necessarily standard normal, then it would be possible to specify a single l_z cut-off value and use this value to make decisions of nonfit. However, the present results indicated that different null distributions, and hence different cut-off values, would be required to evaluate l_z depending on both scale and examinee $\hat{\theta}$ level. This would be extremely awkward to implement in an applied personality assessment situation.

Power of l_z

In contrast to methods used in previous person-fit research, here a nonfitting response was defined as a random response. Because a number of factors appeared to interact in the present study, the detection power results precluded generalizations about conditions that optimize or minimize detection power.

Nevertheless, overall the detection power results were somewhat disconcerting. For example, in the HA and TR scales only slightly over 50% of completely random response vectors were detected at $\alpha = .05$ using ML or EAP scoring. Also, it appeared that an examinee must produce a substantial number of random responses before l_z could reasonably detect nonfit. This observation is especially true when θ was estimated, as it must be in practice. Of course, one remedy to the low detection hit rates would be to evaluate l_z at $\alpha = .1$. Another remedy, given the small number of items in personality measures, may be to combine information across scales and use multitest extensions of the l_z index to assess nonfit (Drasgow, Levine, & McLaughlin, 1991).

The results were based on testing the significance of an observed l_z with respect to a standard normal distribution. To the extent that the standard normal was not the appropriate null distribution, the detection hit rates in Study 2 were biased. Because it was known from Study 1 that the correct null distributions for l_z tended to have variances less than 1.0, the Study 2 results most likely underestimated l_z's ability to detect nonfit. This is just one example in which the standardization problems uncovered in Study 1 clouded interpretation of l_z.

Further insight into the Study 2 hit rate results can be gained by considering various definitions of nonfit. Conceptually, a nonfitting response can be defined as one that provides no psychometric information regarding examinee θ (Reise & Due, 1991). This definition contrasts with the statistical definition, which states that a nonfitting response is one that is low in probability given the model. Because l_z operates under the statistical definition, and the nonfit simulated in this research operated under the former, the present simulations generated

responses that were not necessarily nonfitting by the l_z definition. Hence, a rather severe test of detection power was presented to l_z, resulting in limited detection rates when the number of nonfitting responses was small.

A potentially important finding in the detection power analyses was that hit rates tended to be maximum when the examinee's θ was distant from the average within-scale item difficulty for the TRUE and BIW scoring conditions. Thus, it may be advantageous to administer items that are extreme in difficulty relative to the examinee's estimated θ. It makes sense that the further an item is in difficulty from an examinee's $\hat{\theta}$, the more nonfitting (i.e., the less probable) the nonfitting response will be. Clearly from the l_z equations, items distant in difficulty influence l_z disproportionately as compared to items near the examinee's $\hat{\theta}$.

Previous studies (e.g., Gafni, 1988; Reise & Due, 1991; Schmitt, Cortina, & Whitney, 1993) of person fit have relied heavily on ML estimates of θ and then determining the fit of a response pattern with respect to this estimate. The Study 2 data indicated that scoring method did make a difference. However, the specific effects appeared to be moderated by the scale, examinee θ level, and the number of nonfitting responses. For example, for the AB scale, the difference between hit rates under the three estimated θ conditions and under the TRUE conditions was not great. In contrast, for the HA scale the same discrepancy between θ and $\hat{\theta}$ conditions was substantial.

As a further example, as the number of nonfitting responses increased, the differential effects of the scoring methods increased. In fact, the differences in hit rates between the scoring methods was practically inconsequential when the number of non-fitting responses was small. When the number of nonfitting responses was large, BIW scoring improved hit rates substantially relative to EAP and ML. This result indicates that the differential effects of scoring method interact with the number of nonfitting responses.

Conclusions

Although the results were complex, some general conclusions can be drawn. First, when θ was estimated, regardless of the method used, hit rates were reduced with respect to the TRUE scoring condition. Of course, in practice θ is never known and some estimate must be used to determine person fit. This leads to a paradox. If the θ estimate is "good" (i.e., close to true θ), then nonfit is more detectable. Yet if the estimate is good, there is no reason to be concerned with nonfit. On the other hand, if the estimate is poor, detection power may be lowered. But it is just this circumstance in which it would be most desirable to detect nonfit in order to reject the $\hat{\theta}$ as invalid.

BIW in general appeared to provide superior detection hit rates. Again, however, the benefit of BIW depended on the scale. For instance, BIW was most beneficial in the HA scale and the least in the AB scale, perhaps because of the relative spread of the b parameters: for HA SD(b) = .53 and for AB SD(b) = .89 (Table 1). The implication of these results is that robust scoring may be valuable in terms of identifying lack of person fit. Also, in general there appeared to be little difference in the detection hit rates between ML and EAP scoring methods. One possible explanation of the advantages of BIW scoring was that it might have provided a better estimate of θ than either ML or EAP under the nonfit conditions imposed in Study 2.

References

Birenbaum, M. (1986). Effect of dissimulation motivation and anxiety on response pattern appropriateness measures. *Applied Psychological Measurement, 10,* 167–174.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6,* 431–444.

Drasgow, F. (1982). Choice of test model for appropriateness measurement. *Applied Psychological Measurement, 6,* 297–308.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement, 11,* 59–79.

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

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous

item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38,* 67–86.

Gafni, N. (1988). Detection of systematic and unsystematic aberrancy as a function of ability estimate by several person-fit indices. (Doctoral dissertation, University of Minnesota, 1987). *Dissertation Abstracts International, 49,* 943.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park CA: Sage.

Harnisch, D. L., & Tatsuoka, K. K. (1983). A comparison of appropriateness indices based on item response theory. In R. Hambleton (Ed.), *Applications of item response theory* (pp. 104–122). Vancouver: Educational Research Institute of British Columbia.

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

Mislevy, R. J., & Bock, D. R. (1982). Biweight estimates of latent ability. *Educational and Psychological Measurement, 42,* 725–737.

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

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

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

Schmitt, N., Cortina, J. M., & Whitney, D. J. (1993). Appropriateness fit and criterion-related validity. *Applied Psychological Measurement, 17,* 143–150.

Tellegen, A. (1982). *Brief manual of the Multidimensional Personality Questionnaire.* Unpublished manuscript.

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 CN: JAI Press.

Thissen, D. (1986). *MULTILOG* [Computer program]. Mooresville IN: Scientific Software.

Wainer, H., & Thissen, D. (1987). Estimating ability with the wrong model. *Journal of Educational Statistics, 12,* 339–368.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6,* 473–492.

Author's Address

Send requests for reprints or further information to Steven P. Reise, Department of Psychology, University of California, Riverside CA 92521, U.S.A. Internet: steve@ucrac1.ucr.edu.