The Distribution of Person Fit Using True and Estimated Person Parameters

Michael L. Nering
University of Minnesota

A variety of methods have been developed to determine the extent to which a person's response vector fits an item response theory model. These person-fit methods are statistical methods that allow researchers to identify nonfitting response vectors. The most promising method has been the l_z statistic, which is a standardized person-fit index. Reise & Due (1991) concluded that under the null condition (i.e., when data were simulated to fit the model) l_z performed reasonably well. The present study extended the findings of past researchers (e.g., Drasgow, Levine, & McLaughlin, 1987; Molenaar & Hoijtink, 1990; Reise and Due). Results show that l_z may not perform as expected when estimated person parameters ($\hat{\theta}$) are used rather than true θ. This study also examined the influence of the pseudo-guessing parameter, the method used to identify nonfitting response vectors, and the method used to estimate θ. When θ was better estimated, l_z was more normally distributed, and the false positive rate for a single cut score did not characterize the distribution of l_z. Changing the c parameter from .20 to 0.0 did not improve the normality of the l_z distribution. *Index terms: appropriateness measurement, Bayesian estimation, item response theory, maximum likelihood estimation, person fit.*

During the past 15 years, several methods have been proposed to investigate a person's response pattern to a set of test items to determine whether that person is being accurately measured in the context of item response theory (IRT) models (e.g., Levine & Rubin, 1979; Tatsuoka & Linn, 1983; Trabin & Weiss, 1983). Known in the past as appropriateness measurement (e.g., Drasgow, Levine, & Williams, 1985; Tatsuoka, 1984), this area of research is now commonly referred to as person fit (e.g., Reise, 1990; Reise & Due, 1991).

Many methods for indexing person fit make use of the three-parameter logistic model (3PLM):

$$P(\theta_j) = c_i + (1 - c_i)\frac{\exp\left[Da_i(\theta_j - b_i)\right]}{1 + \exp\left[Da_i(\theta_j - b_i)\right]}, \tag{1}$$

where

i indexes the item,
j indexes the person,
c_i is the pseudo-guessing parameter,
a_i is the item discrimination parameter,
b_i is the item difficulty parameter,
D is a scaling constant equal to approximately 1.7, and
θ_j is the person location parameter for person j.

The standardized version of the l_o index developed by Levine & Rubin (1979) has been studied by person-fit researchers (e.g., Reise & Due, 1991; Schmitt, Cortina, & Whitney, 1993). The l_o index was the first to use the maximum of the likelihood function for indexing person fit. This index represents the log of the peak of the likelihood function, which can be found by multiplying the probabilities of the correct and incorrect responses to the items calculated by Equation 1. l_o is defined as

$$l_o = \ln \prod_{i} P_i(\hat{\theta})^{u_i} Q_i(\hat{\theta})^{(1-u_i)}, \tag{2}$$

where

P_i refers to the probability of a person correctly answering item i,
Q_i refers to the probability of a person incorrectly answering the item,
u_i is the scored (1 or 0) response to item i, and

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 19, No. 2, June 1995, pp. 121–129
© Copyright 1995 Applied Psychological Measurement Inc.
0146-6216/95/020121-09$1.70

$\hat{\theta}$ is the maximum likelihood estimate of θ.

The extent to which a given response vector fits the IRT model is reflected by the degree to which the likelihood function is peaked. If a response pattern overfits the model (e.g., a Guttman vector), then the likelihood function will be very peaked, relative to that of other response vectors. If a response pattern does not fit the model (e.g., a reverse Guttman vector), then the peak of the function will be very low. Drasgow et al. (1985) found that the l_o distribution changes as a function of θ. Thus, individuals at one position on the θ continuum might be more likely to be identified as not fitting the IRT model than individuals at another position on the continuum. By standardizing l_o, Drasgow et al. (1985) overcame this interaction between θ and person fit. The standardized index, referred to here as l_z, can be defined as

$$l_z = \frac{l_o - \mathrm{E}(l_o)}{\left[\mathrm{Var}(l_o)\right]^{1/2}} , \tag{3}$$

where, for a given value of $\hat{\theta}$, $\mathrm{E}(l_o)$ is the expected value of l_o,

$$\mathrm{E}(l_o) = \sum_{i=1}^{n} \left\{ P_i(\hat{\theta}) \ln P_i(\hat{\theta}) + \left[1 - P_i(\hat{\theta})\right] \ln\left[1 - P_i(\hat{\theta})\right] \right\}, \tag{4}$$

and $\mathrm{Var}(l_o)$ is the variance of l_o,

$$\mathrm{Var}(l_o) = \sum_{i=1}^{n} P_i(\hat{\theta})\left[1 - P_i(\hat{\theta})\right] \left\{ \ln \left\{ \frac{P_i(\hat{\theta})}{\left[1 - P_i(\hat{\theta})\right]} \right\} \right\}^2 . \tag{5}$$

The distribution of l_z, as proposed by Drasgow et al. (1985), uses asymptotic theory. Thus, as the number of items increases, the l_z distribution approaches a standard distribution with a mean of 0.0 and a standard deviation (SD) of 1.0. In addition, because l_z represents the sum of independent random variables, it should approach a normal distribution as the number of items increases. Thus, an individual having an l_z value of −1.99 would be considered nonfitting, because their l_z value represents a value beyond the .05 error value of −1.96 for a two-tailed test.

Previous researchers (e.g., Drasgow et al., 1987; Molenaar & Hoijtink, 1990; Reise & Due, 1991) have used the proportion of nonfitting individuals in a sample to determine the extent to which the l_z statistic is normally distributed. Reise and Due concluded that under the null condition (i.e., when data were simulated to fit the IRT model) the l_z statistic performed reasonably well. They also concluded that when nonfitting response patterns were simulated, the spread of the b and c parameters and test length all influenced the l_z statistic; as the variability in the item difficulties increased, the c parameter decreased, and the test length increased, the ability to accurately identify nonfitting response vectors became less difficult. Although previous research used much longer tests to investigate person fit [e.g., Drasgow (1982) used a 95-item test], Reise and Due found that l_z performed satisfactorily with 21 items.

Previous person-fit research has had several problems. For example, Reise & Due (1991) used true θ parameters in their simulations to calculate l_z. However, Equations 2–5 show that l_z is conditional on $\hat{\theta}$. Person-fit researchers have commonly used the proportion of simulees having an l_z value less than a criterion value to evaluate the performance of the index (e.g., Drasgow et al., 1987; Molenaar & Hoijtink, 1990; Reise and Due). This assumes that l_z is normally distributed. Finally, nonfitting response patterns have been simulated in ways that might not necessarily be found in real data. For example, to simulate different levels of person fit, Reise and Due included a person discrimination parameter (a_p) in the equation for the 3PLM:

$$P(\theta) = c_i + (1 - c_i) \frac{\exp\left[Da_i a_p(\theta - b_i)\right]}{1 + \exp\left[Da_i a_p(\theta - b_i)\right]} . \tag{6}$$

The rationale for using this additional person parameter to simulate nonfitting responses developed from an area of research known as person response theory (PRT; e.g., Kiely, 1992; Strandmark & Linn, 1987). In PRT, if an item is less discriminating for a particular person, then that person's θ is not being accurately measured and the $\hat{\theta}$ is inappropriate. Because a_i is multiplied by a_p, as an item becomes less discriminating for a person (i.e., when a_p becomes less than 1.0), the slope of the item response function (IRF) will decrease. This decrease in the slope of the IRF will be reflected in the peak of the likelihood function, because the slope of the IRF is inversely related to the height of the peak of the likelihood function (Hambleton & Swaminathan, 1985). Reise & Due (1991) used a_p to generate their data, and then calculated l_z without the a_p parameter (effectively setting $a_p = 1.0$). The problem with this approach is that it may not be an accurate way of simulating nonfitting response patterns that might be found in a dataset that has not been simulated.

The first goal of the present investigation was to demonstrate that l_z does not necessarily follow a normal distribution when θ is estimated. Datasets were simulated (using monte carlo methods) to fit the 3PLM, and then person-fit indexes were calculated for both θ and $\hat{\theta}$. A second goal was to determine whether the method used to estimate θ (maximum likelihood or Bayesian) had an influence on the l_z statistic. A third goal was to investigate whether the −1.64 criterion used by Reise & Due (1991) is an appropriate method for identifying nonfitting persons when θ is estimated. The final goal was to evaluate l_z within the context of a short test (i.e., 25 items). This was done so that comparisons could be made to Reise and Due and to investigate the distribution of l_z when asymptotic theory is not satisfied (which is the case with many tests).

Method

Design

The present study investigated the performance of l_z as a function of variations in the θ distribution, the spread of the b parameters, the method used to estimate θ, and the value of the c parameter. l_z was calculated for all conditions presented in Table 1. The item by person matrices and the item parameters were simulated, using monte carlo methods, to fit the 3PLM (Yoes, 1993).

For all datasets, the item discrimination was fixed at 1.5, the number of simulees was 1,000, and the number of items was 25. For Conditions 1–3, θ was fixed at 0.0 (see Table 1); for Conditions 4–6, θ was distributed approximately as N(0.0,.5); and for Conditions 7–9, θ was approximately N(0.0, 1.0). Thus, there were three different θ distributions.

Table 1
Distributions of θ, b, and c, By Condition

Condition	Distribution of θ	Distribution of b	c
True θ			
1	0.0	U(−1.0, 1.0)	.2
2	0.0	U(−2.0, 2.0)	.2
3	0.0	U(−3.0, 3.0)	.2
4	N(0, .5)	U(−1.0, 1.0)	.2
5	N(0, .5)	U(−2.0, 2.0)	.2
6	N(0, .5)	U(−3.0, 3.0)	.2
7	N(0, 1)	U(−1.0, 1.0)	.2
8	N(0, 1)	U(−2.0, 2.0)	.2
9	N(0,1)	U(−3.0, 3.0)	.2
Maximum Likelihood $\hat{\theta}$			
10	0.0	U(−1.0, 1.0)	.2
11	0.0	U(−2.0, 2.0)	.2
12	0.0	U(−3.0, 3.0)	.2
13	N(0, .5)	U(−1.0, 1.0)	.2
14	N(0, .5)	U(−2.0, 2.0)	.2
15	N(0, .5)	U(−3.0, 3.0)	.2
16	N(0, 1)	U(−1.0, 1.0)	.2
17	N(0, 1)	U(−2.0, 2.0)	.2
18	N(0, 1)	U(−3.0, 3.0)	.2
Bayesian $\hat{\theta}$			
19	N(0, 1)	U(−1.0, 1.0)	.2
20	N(0, 1)	U(−2.0, 2.0)	.2
21	N(0, 1)	U(−3.0, 3.0)	.2
True θ			
22	0.0	U(−1.0, 1.0)	0.0
23	0.0	U(−2.0, 2.0)	0.0
24	0.0	U(−3.0, 3.0)	0.0
25	N(0, .5)	U(−1.0, 1.0)	0.0
26	N(0, .5)	U(−2.0, 2.0)	0.0
27	N(0, .5)	U(−3.0, 3.0)	0.0
28	N(0, 1)	U(−1.0, 1.0)	0.0
29	N(0, 1)	U(−2.0, 2.0)	0.0
30	N(0, 1)	U(−3.0, 3.0)	0.0
Maximum Likelihood $\hat{\theta}$			
31	0.0	U(−1.0, 1.0)	0.0
32	0.0	U(−2.0, 2.0)	0.0
33	0.0	U(−3.0, 3.0)	0.0
34	N(0, .5)	U(−1.0, 1.0)	0.0
35	N(0, .5)	U(−2.0, 2.0)	0.0
36	N(0, .5)	U(−3.0, 3.0)	0.0
37	N(0, 1)	U(−1.0, 1.0)	0.0
38	N(0, 1)	U(−2.0, 2.0)	0.0
39	N(0, 1)	U(−3.0, 3.0)	0.0
Bayesian $\hat{\theta}$			
40	N(0, 1)	U(−1.0, 1.0)	0.0
41	N(0, 1)	U(−2.0, 2.0)	0.0
42	N(0, 1)	U(−3.0, 3.0)	0.0

Each θ distribution had three sets of *b* distributions: U(−1.0, +1.0), U(−2.0, +2.0), and U(−3.0, +3.0). This was designed to determine whether the interaction between the spread of the *b* parameter and the distribution of θ affected the normality of the l_z distribution.

The person by item matrices and their associated true item parameter matrices, used in Conditions 1–9, also were used to estimate θ by maximum likelihood estimation (Hambleton & Swaminathan, 1985). Thus, the same datasets that were used in Conditions 1–9 (see Table 1) also were used in Conditions 10–18. In Conditions 19–21, Bayesian $\hat{\theta}$s (see Hambleton & Swaminathan, 1985) were calculated for the datasets that were used in Conditions 7, 8, and 9. This was done only for Conditions 7, 8, and 9 because the Bayesian prior distribution was N(0.0, 1.0) and thus accurately reflected the distribution of θ in these conditions. Weiss & McBride (1984) demonstrated that when an accurate prior distribution is used, there is less bias in the Bayesian estimation of θ than there is for an inaccurate prior distribution.

Table 1 shows that in Conditions 1–21 $c = .2$. Conditions 22–42 were the same as Conditions 1–21 except $c = 0.0$. It was assumed that as *c* increases there is a loss in psychometric information, which should have adverse affects on l_z (Reise & Due, 1991). If the *c* parameter causes the l_z distribution to become different from what is expected, then setting $c = 0.0$ should aid in l_z becoming more normally distributed (A. Due, personal communication, May, 1994).

Evaluation of l_z

The normality of the l_z distribution was evaluated in terms of the mean, SD, skewness, and kurtosis. The Kolmogorov-Smirnov (KS) goodness-of-fit test also was performed. False positives (i.e., the proportion of model-fitting simulees categorized as not fitting the model using the normal curve value for a .05 error rate) also were calculated to determine if they accurately reflected the underlying distribution of l_z. Two false positive calculation methods were used to identify nonfitting persons: FP1 = $l_z \leq -1.64$, and FP2 = $l_z \leq$ [mean − 1.64 (SD)]. The mean and SD used to find FP2 were calculated from the distribution of l_z. FP1 and FP2 were calculated to make comparisons to the findings in Reise & Due (1991) who used FP1; FP2 takes into consideration the mean and SD of the l_z distribution in identifying nonfitting response vectors.

To evaluate the role of the θ estimation process on l_z, Pearson product-moment correlations, root mean square error (RMSE), and average signed bias (ASB) were calculated between θ and $\hat{\theta}$. RMSE and ASB can be defined, respectively, as

$$\text{RSME} = \exp\left[\frac{\sum_{j=1}^{N}(\theta_j - \hat{\theta}_j)^2}{N}\right]^{1/2}, \tag{7}$$

and

$$\text{ASB} = \frac{\sum_{j=1}^{N}(\theta_j - \hat{\theta}_j)}{N}, \tag{8}$$

where *j* represents the number of simulees ($j = 1, \ldots, 1{,}000$).

Results

Estimation of θ

The RMSE, ASB, and the correlation between θ and $\hat{\theta}$ [$\rho(\theta, \hat{\theta})$] are reported in Table 2. In Conditions 16 and 17, the estimates of θ were very different from the θ values. The RMSE in these two conditions were .666 and 1.436, respectively; the ASB values were .308 and 1.242, respectively. The RMSEs in the remaining conditions in Table 2 were consistently above .2, and the ASBs were very close to 0.0. As with l_z, the estimation of θ became poorer when the spread of the *b*s was small relative to the SD of the θ distribution.

$\rho(\theta, \hat{\theta})$ in Table 2 were all above .77, and above .92 when a Bayesian prior was used to estimate θ. Although the simulees had approximately the same rank order when θ was estimated, the RMSE suggests that there is estimation error, and the ASB suggests that the error did not have a consistent positive or a consistent negative bias across the various conditions studied.

Table 2
RMSE, ASB, and Correlation Between θ and $\hat{\theta}$ [$\rho(\theta, \hat{\theta})$] for Estimation Conditions

Condition	RMSE	ASB	$\rho(\theta, \hat{\theta})$
Maximum Likelihood $\hat{\theta}$			
10	.255	–.014	[a]
11	.361	–.032	[a]
12	.396	–.049	[a]
13	.281	–.009	.872
14	.369	–.017	.804
15	.418	–.053	.775
16	.666	.308	.843
17	1.436	1.242	.783
18	.339	–.032	.924
Bayesian $\hat{\theta}$			
19	.331	–.012	.944
20	.332	–.013	.942
21	.366	–.011	.928
Maximum Likelihood $\hat{\theta}$			
31	.202	–.007	[a]
32	.254	.003	[a]
33	.269	.014	[a]
34	.215	.007	.813
35	.251	.001	.899
36	.268	.006	.881
37	.243	–.004	.957
38	.271	.010	.957
39	.298	–.019	.953
Bayesian $\hat{\theta}$			
40	.277	–.003	.959
41	.267	.007	.961
42	.280	–.004	.957

[a] $\rho(\theta, \hat{\theta})$ could not be computed because θ was fixed.

The RMSE, the ASB, and $\rho(\theta, \hat{\theta})$ were very similar when $c = 0.0$. The largest reduction in RMSE values occurred for Conditions 37 and 38 as compared to Conditions 16 and 17 (.666 versus .243 and 1.436 versus .271, respectively). The ASB values were generally closer to 0.0 when $c = 0.0$ (Conditions 31–42) than when $c = .2$ (Conditions 10–21). For the $c = 0.0$ conditions, Condition 31 had the smallest RMSE of .202. Thus, setting $c = 0.0$ did not appear to entirely alleviate (but it reduced) the estimation error that was found in Conditions 10–21, except when the relationship between the variability associated with θ was greater than the variability associated with the b distribution (such as in Conditions 16 and 17). $\rho(\theta, \hat{\theta})$ did not seem to be affected by the estimation procedure when $c =$ 0.0; however, it was when $c = .2$. In Conditions 16 and 17, when ML estimation was used, the correlations were .843 and .783, respectively; these values increased to .944 and .924 in Conditions 19 and 20 when Bayesian estimation was used. Very little change in the correlations was found in the same conditions when $c = 0.0$ (i.e., comparing Conditions 37 and 38 to Conditions 40 and 41, respectively).

The high correlations in Table 2 may suggest that because θ and $\hat{\theta}$ are highly correlated, l_z cannot vary as a function of $\hat{\theta}$ being different from θ. However, these correlations only provide information about the relative rank ordering of θ and $\hat{\theta}$. The RMSE values suggest that there are differences between θ and $\hat{\theta}$ that are not reflected in the correlations. As a consequence, different values of l_z may result from the misestimation of θ.

Distribution of l_z

Table 3 shows that for Conditions 1–9 l_z based on true θ had means and SDs that were very close to their predicted (0.0, 1.0) values. The mean for Condition 7 (.05) was the most different from 0.0. The SD for Condition 9 (.971) was the most different from 1.0. The distribution of l_z was significantly negatively skewed in all of these conditions and in most conditions was leptokurtic. Only in Conditions 1 and 7 were the kurtosis indexes not significantly leptokurtic. The results of the KS tests suggest that in Conditions 1–9 l_z consistently did not follow a normal distribution.

Results differed considerably when θ was estimated. In Conditions 16 and 17 (i.e., when the SD of θ was greater than the spread of the bs), the means and SDs [(–.652, 1.49) and (–1.33, .971), respectively] clearly demonstrated that l_z was not standardized. In Conditions 10–15 and Conditions 18–21, the mean was consistently above 0.0, and the SD of l_z was less than 1.0. The average mean l_z value across these 10 conditions was approximately .18, and the average SD was approximately .90. The degree of kurtosis found in Conditions 10–21 was not systematic; only when $\theta \sim N(0.0, .5)$ (Conditions 13–15) were the indexes significant regardless of the spread of the items. The skewness of the l_z distribution in Condition 17 was –.078, which was the only l_z dis-

Table 3
Mean, Standard Deviation (SD), Kurtosis, Skewness, Kolmogorov-Smirnov (KS) Statistic, and Sample Size (N) for l_z by Condition

Condition	Mean	SD	Kurtosis	Skewness	KS	N^a
True θ						
1	.019	.974	.153	−.606**	1.623*	1,000
2	−.014	.996	.344*	−.613**	1.499*	1,000
3	.013	1.002	1.598**	−.898**	1.957***	1,000
4	.014	.990	.459**	−.636**	1.511*	1,000
5	−.025	1.021	.963**	−.793**	1.788**	1,000
6	−.014	1.035	.908**	−.885**	2.038***	1,000
7	.050	1.011	.291	−.565**	1.978***	1,000
8	.001	1.017	1.420**	−.862**	2.352***	1,000
9	.008	.971	.572**	−.692**	1.758**	1,000
Maximum Likelihood $\hat{\theta}$						
10	.153	.826	−.047	−.398**	1.283	1,000
11	.200	.941	.267	−.574**	1.718**	1,000
12	.172	.887	.854**	−.665**	1.914**	1,000
13	.139	.818	.371**	−.368**	1.216	996
14	.188	.926	.673**	−.703**	1.904**	1,000
15	.174	.907	1.086**	−.836**	1.951**	1,000
16	−.652	1.492	1.824**	−1.033**	1.675**	773
17	−1.332	.971	.056	−.078	.591	945
18	.188	.876	.264	−.605**	2.085***	994
Bayesian $\hat{\theta}$						
19	.263	.810	.243	−.434**	2.081***	1,000
20	.197	.908	1.175**	−.853**	2.352***	1,000
21	.185	.878	.400**	−.661**	1.939**	1,000
True θ						
22	−.026	1.057	.306*	−.547**	1.578*	1,000
23	.050	.969	.716**	−.848**	2.226***	1,000
24	−.022	1.008	1.819**	−1.256**	3.435***	1,000
25	−.029	1.000	.187	−.564**	1.389*	1,000
26	−.034	.985	.970**	−.880**	2.093***	1,000
27	.017	1.001	1.547**	−1.140**	2.708***	1,000
28	.011	.999	1.921**	−1.059**	2.268***	1,000
29	.011	.983	1.718**	−1.083**	2.471***	1,000
30	.035	.958	1.273**	−1.034**	2.764***	1,000
Maximum Likelihood $\hat{\theta}$						
31	.112	1.056	.373*	−.587**	1.614*	1,000
32	.239	.955	1.094**	−1.017**	2.952***	1,000
33	.151	.931	1.826**	−1.262**	3.085***	1,000
34	.068	.916	.374*	−.473**	1.425*	992
35	.134	.962	1.101**	−.937**	2.281***	1,000
36	.203	.931	1.868**	−1.264**	3.280***	1,000
37	.094	.818	1.749**	−.751**	2.569***	918
38	.182	.954	2.035**	−1.152**	2.543***	978
39	.228	.885	2.065**	−1.227**	3.057***	990
Bayesian $\hat{\theta}$						
40	.253	.806	2.138**	−1.099**	3.310***	1,000
41	.238	.944	2.220**	−1.228**	2.739***	1,000
42	.268	.874	1.918**	−1.215**	3.078***	1,000

Note. When $N < 1{,}000$, maximum likelihood estimates could not be calculated because some simulees had a response pattern of all 0s or 1s.
$^*p < .05$; $^{**}p < .01$; $^{***}p < .001$.

tribution among the 42 conditions not significantly skewed (see Table 3). According to the KS test, l_z did follow a normal distribution for Conditions 10, 13, and 17; however, these were the only conditions among the 42 studied in which the distributions were considered normal.

Comparing Conditions 10–21 to Conditions 22–42, the c parameter appears to have negligible influence on the mean and SD of l_z. The kurtosis and skewness of l_z, however, tended to increase as c decreased from .2 to 0.0.

The means and SDs of l_z were very close to the hypothesized values when θ was used and $c = 0.0$ (Conditions 22–30). The mean for Condition 23 (.05) was the most different from 0.0 for these conditions, and the SD for Condition 30 (.985) was the most different from 1.0. For θ and $c = 0.0$, Condition 25 was the only distribution of l_z that was not significantly leptokurtic. All indexes of skewness in Conditions 22–30 were statistically significant.

In Conditions 31–39, the mean of l_z was consistently above 0.0. Condition 34 had the smallest mean (.068) for this set of conditions; however, the SD was .916. The SDs of l_z when θ was estimated using maximum likelihood were consistently below 1.0, except for Condition 31 (1.056). All of the l_z distributions in Conditions 31–39 were significantly leptokurtic, and negatively skewed. Bayesian estimation of θ had little effect on l_z or the detection of misfit—results in Tables 3 and 4 differed only slightly between the ML conditions (Conditions 16–18 and 28–30) and the comparable Bayesian conditions (19–21 and 40–42, respectively).

False Positives

The proportion of false positives was relatively consistent in Conditions 1–9 for both FP1 and FP2 (see Table 4). The false positive rates were all above .05 (the expected value), and they were very similar to the values found by Reise & Due (1991). FP1 was more inconsistent when $\hat{\theta}$ was used (Conditions 10–21 and 31–42). In Conditions 16 and 17, the proportion of simulees identified as not fitting the model by FP1 was .212 and .367, respectively. In Conditions 10–15 and 18–21, the proportion of

Table 4
FP1 and FP2 for θ and $\hat{\theta}$ Values

Condition	FP1	FP2
True θ		
1	.060	.071
2	.067	.067
3	.061	.061
4	.058	.061
5	.073	.070
6	.079	.069
7	.063	.068
8	.070	.068
9	.057	.062
Maximum Likelihood $\hat{\theta}$		
10	.024	.066
11	.034	.065
12	.028	.053
13	.025	.059
14	.045	.067
15	.040	.063
16	.212	.065
17	.367	.055
18	.032	.057
Bayesian $\hat{\theta}$		
19	.022	.065
20	.039	.063
21	.029	.058
True θ		
22	.074	.061
23	.055	.068
24	.069	.064
25	.067	.064
26	.069	.066
27	.069	.071
28	.061	.061
29	.060	.069
30	.061	.069
Maximum Likelihood $\hat{\theta}$		
31	.060	.061
32	.042	.069
33	.050	.064
34	.044	.054
35	.054	.070
36	.049	.075
37	.029	.064
38	.048	.067
39	.041	.074
Bayesian $\hat{\theta}$		
40	.023	.072
41	.042	.072
42	.040	.073

simulees identified as not fitting the model was less than .05, ranging from .022 to .045. The proportion of individuals identified as not fitting using FP2 was very close to .05 in all conditions in Table 4, indicating that estimation error did not affect FP2 as it did FP1.

The false positive method (FP1 and FP2) did not appear to change the proportion of simulees identified as not fitting the model when l_z was calculated using θ and $c = 0.0$. As shown in Table 4, the false positives in Conditions 22–30 were between .055 and .074 regardless of the false positive method. In Conditions 31–42 the false positives ranged from .023 to .060 for FP1, and thus were more consistent than the FP1 values obtained for Conditions 10–21.

Discussion

One method that has been used to investigate the performance of l_z under a variety of conditions is the proportion of examinees identified as not fitting the model (Drasgow et al., 1987; Molenaar & Hoijtink, 1990; Reise & Due, 1991). This assumes that l_z has an underlying normal distribution. As demonstrated by the indexes of skewness and kurtosis in Table 3, this assumption of normality should be questioned when θ is estimated. Because the mean and SD of l_z were consistently different from 0.0 and 1.0 in this study when $\hat{\theta}$ was used (see Table 3), the interpretability of l_z may be more complicated than believed by past researchers (e.g., Reise and Due).

False positives were more consistent when FP2 was used. In Table 4, FP2 values for Conditions 10–21 were very similar to what was found by Reise & Due (1991), who used FP1. However, when $c = 0.0$ the false positive method became less important. When using the 3PLM, the false positive method should be considered carefully. The results of this study suggest that the false positive rate for a single cut score may not provide a completely adequate characterization of l_z, because the underlying distribution of the statistic is not taken into consideration.

As indicated above, it might be expected that when $c = 0.0$ l_z would be more normally distributed than when $c > 0.0$. Under these circumstances, θ was better estimated (see Table 2). However, the mean and SD of l_z were more consistent with $c = 0.0$, and the kurtosis and skewness increased as c changed from .2 to 0.0. Thus, setting $c = 0.0$ did not improve the normality of the l_z distribution.

Better estimation of θ should improve the estimation of person fit. This study demonstrated that when the prior θ distribution matched the true θ distribution (the Bayesian estimation conditions), the estimation of θ and l_z improved compared to the conditions in which θ was estimated using maximum likelihood. By setting $c = 0.0$ and using Bayesian estimation of θ (with a prior distribution that matched the true θ distribution), l_z had a mean and a SD that were closer to the expected values, and the estimation of θ improved. However, in real data situations, the true prior distribution is unknown.

In many measurement situations, the tests are of such a length that they do not satisfy asymptotic theory. The tests used here consisted of only 25 items. For tests that do not satisfy asymptotic theory, the results presented in Table 3 suggest that in many cases the distribution of l_z will not follow a standard normal distribution. Although some of the distributions investigated here only moderately differed from standard normal, this departure can affect the interpretation of l_z.

In this study, only true item parameters were used; future research should investigate the influence of estimated item parameters on l_z. Because FP2 is a sample-dependent method of identifying nonfitting persons, future research should also be concerned with identifying a cutoff value that is appropriate across a variety of testing conditions. Test length was not varied here. With continuing advancements being made in test administration (e.g., computerized adaptive testing) future research should also investigate the influence of test length on the distribution of person-fit statistics.

References

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

Drasgow, F., Levine, M. L., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychologi-*

cal Measurement, 11, 59–79.

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.

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

Kiely, G. L. (1992). *The robustness of item and person parameter estimates to variation in person discrimination in the two-parameter logistic model* (Doctoral dissertation, University of Minnesota, 1992). *Dissertation Abstracts International, 53,* 08B.

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

Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person fit. *Psychometrika, 55,* 75–106.

Reise, S. P. (1990). A comparison of item- and person-fit methods of assessing model-data fit in IRT. *Applied Psychological Measurement, 14,* 127–137.

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

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

Strandmark, N. L., & Linn, R. L. (1987). A generalized logistic item response model parameterizing test score inappropriateness. *Applied Psychological Measurement, 11,* 355–370.

Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika, 49,* 95–110.

Tatsuoka, K. K., & Linn, R. L. (1983). Indices for detecting unusual patterns: Links between two general approaches and potential applications. *Applied Psychological Measurement, 7,* 81–96.

Trabin, T. E., & Weiss, D. J. (1983). The person response curve: Fit of individuals to item characteristic curve models. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 83–108). New York: Academic Press.

Weiss, D. J., & McBride, J. R. (1984). Bias and information of Bayesian adaptive testing. *Applied Psychological Measurement, 8,* 273–285.

Yoes, M. E. (1993). *A comparison of the effectiveness of item parameter estimation techniques used with the 3-parameter logistic item response theory model* (Doctoral dissertation, University of Minnesota, 1993). *Dissertation Abstracts International, 54,* 12B.

Acknowledgments

The author expresses his gratitude to an anonymous reviewer for his/her insightful comments on an earlier draft of this paper.

Author's Address

Send requests for reprints or further information to Michael L. Nering, Department of Psychology, University of Minnesota, 75 East River Road, Minneapolis MN 55455, U.S.A. Internet: neri0001@gold.tc.umn.edu.