Methodology review

Meijer, R.R.; Sijtsma, K.

Published in:
Applied Psychological Measurement

DOI:
10.1177/01466210122031957

Publication date:
2001

Document Version
Publisher's PDF, also known as Version of record

Methodology Review: Evaluating Person Fit

Rob R. Meijer and Klaas Sijtsma

Applied Psychological Measurement 2001; 25; 107
DOI: 10.1177/01466210122031957

Methodology Review: Evaluating Person Fit

Rob R. Meijer, University of Twente

Klaas Sijtsma, Tilburg University

Person-fit methods based on classical test theory and item response theory (IRT), and methods investigating particular types of response behavior on tests, are examined. Similarities and differences among person-fit methods and their advantages and disadvantages are discussed. Sound person-fit methods have been derived for the Rasch model. For other IRT models, the empirical and theoretical distributions differ for most person-fit statistics when used with short and moderate length tests. The detection rate of person-fit statistics depends on the type of misfitting item-score patterns, test length, and trait levels. The usefulness of person-fit statistics for improving measurement depends on the application.
Index terms: appropriateness measurement, caution indices, item response theory, person fit, test theory.

Since the beginning of standardized testing, measurement inaccuracy has received widespread attention. Attempts to understand and mitigate measurement inaccuracy have been based on reliability theory and methods for estimating reliability (Gulliksen, 1950; Lord & Novick, 1968; Spearman, 1910), statistics comparing groups with respect to the probability of correctly answering an item (differential item functioning; e.g., Millsap & Everson, 1993), differential identification of person subgroups using multiple regression analysis with dummy variables (e.g., Aguinis & Stone-Romero, 1997), and—the focus of this review—methods for determining the fit of individual item-score patterns to a test model.

Information that supplements a test score can be obtained by studying patterns of individual item scores (e.g., Nunnally, 1978, p. 437). Discriminant analysis and cluster analysis have been used for clustering similar types of score patterns (Nunnally, 1978, pp. 453–467) to determine whether subgroups can be distinguished. Both methods, however, focus on groups, not on individual persons. In this review, methods are reviewed that provide information at the individual level and that detect misfitting item-score patterns. Overviews of these methods were given by Hulin, Drasgow, & Parsons (1983, Ch. 4), Kogut (1986), and Meijer & Sijtsma (1995).

Person Fit or Appropriateness Measurement

Methods evaluating the fit of an individual's test performance to an item response theory (IRT) model have been referred to as "appropriateness measurement," or more recently, "person-fit" methods. Levine & Drasgow (1983) limited appropriateness measurement methods to those that "recognize inappropriate test scores" (p. 110). However, in most appropriateness measurement and person-fit studies, response behavior is described on the basis of some type of test model. This implies that the appropriateness of a test score is defined by the fit of an item-score pattern to a test model. Person-fit methods here refer to statistical methods for evaluating the misfit of individual test performance to an IRT model or other item-score patterns in a sample of persons.

Rationale for Person-Fit Research

The number-correct (NC) score (or a trait level estimate) might be inadequate as a measure of a person's trait level. For example, a person could guess the correct answers to multiple-choice

Applied Psychological Measurement, Vol. 25 No. 2, June 2001, 107–135

items, thus raising his/her score on a test; or, a person not familiar with the test format could, due to this unfamiliarity, obtain a lower score than expected (e.g., Wright & Stone, 1979, pp. 165–190). Inaccurate measurement of the trait level also might be caused by: "sleeping" behavior (e.g., inaccurately answering the first questions in a test because of problems getting started), cheating (e.g., copying answers from another examinee), and "plodding" behavior (e.g., working very slowly and methodically, thus generating item-score patterns that are overly ideal, given the stochastic nature of response behaviors as assumed by most IRT models; see Ellis & van den Wollenberg, 1993; Holland, 1990).

Not all types of unusual behavior affect test scores. For example, a person might guess correctly on some items and incorrectly on others. Due to the stochastic nature of guessing, this might not result in substantially different test scores under most IRT models. Whether aberrant behavior leads to misfitting item-score patterns depends on numerous factors, such as the type and amount of aberrant behavior.

All methods discussed here can be used to detect misfitting item-score patterns. However, several of these methods do not allow the recovery of the mechanism that created the deviant item-score patterns. Other methods explicitly test against specific violations of a test model assumption or particular types of deviant item-score patterns. The latter group of methods could facilitate the interpretation of misfitting item-score patterns. Most person-fit statistics compare a person's observed and expected item scores across test items. Expected item scores are determined on the basis of either an IRT model or observed item data in a sample of persons.

Table 1 gives an overview of the statistics discussed here, categorized according to the model for which they were developed. This categorization should not be interpreted too strictly. For example, the U statistic, developed for the Rasch model, also can be used for other IRT models.

Table 1
Person-Fit Statistics

Group-Based and Nonparametric	2PLM and 3PLM
r_{pbis}, r_{bis} (Donlon & Fischer, 1968)	l_0 (Levine & Rubin, 1979)
C (Sato, 1975)	D (Weiss, 1973; Trabin & Weiss, 1983)
U (van der Flier, 1980; Meijer, 1994)	*ECI* statistics (Tatsuoka, 1984)
A_i, D_i, E_i (Kane & Brennan, 1980)	l_z (Drasgow, Levine, & Williams, 1985)
C^* (Harnisch & Linn, 1981)	*JK*, *O/E* (Drasgow, Levine, & McLaughlin, 1987)
ZU3 (van der Flier, 1982)	l_{zm} (Drasgow, Levine, & McLaughlin, 1991)
NCI, *ICI* (Tatsuoka & Tatsuoka, 1983)	c (Levine & Drasgow, 1988)
H_i^T (Sijtsma, 1988; Sijtsma & Meijer, 1992)	
Rasch Models	CAT
U (Wright & Stone, 1979)	K (Bradlow, Weiss, & Cho, 1998)
W (Wright & Masters, 1982)	T statistics (van Krimpen-Stoop & Meijer, 2000)
UB, *UW* (Smith, 1985)	Z_c (McLeod & Lewis, 1999)
M (Molenaar & Hoijtink, 1990)	
χ^2_{sc} (Klauer & Rettig, 1990)	
$T(X)$ (Klauer, 1991, 1995)	

Person-Fit Methods Based on Group Characteristics

Statistics

A general formula is used to demonstrate similarities among person-fit statistics. Let a particular choice of weights (w) define a particular person-fit statistic. Assume that n examinees take a test of k items. Let π_g denote the proportion-correct score on item g, estimated by $\hat{\pi}_g = n_g/k$, where n_g

is the number of items scored 1. Let the items be ordered and numbered according to a decreasing proportion-correct score (increasing item difficulty): $\pi_1 > \pi_2 > \ldots > \pi_k$. Let the realization of a dichotomous (0, 1) item score be denoted $X_g = x_g$ $(g = 1, 2, \ldots, k)$. Examinees are indexed by i, with $i = 1, 2, \ldots, n$. The NC score $X = r$ is the unweighted sum of item scores ($\sum_{g=1}^{k} X_g = r$).

Most group-based person-fit statistics compare a count of certain score patterns for item pairs with the expectation under the deterministic Guttman (1944, 1950) model:

$$\theta < \delta_g \Leftrightarrow P_g(\theta) = 0 \,, \tag{1}$$

and

$$\theta \geq \delta_g \Leftrightarrow P_g(\theta) = 1 \,, \tag{2}$$

where

θ is the person's latent trait level,

δ is an item location parameter, which is a value on the θ scale, and

$P_g(\theta)$ is the conditional probability of a correct answer to item g.

The Guttman model excludes a correct answer on a relatively difficult item h, and an incorrect answer on an easier item g, by the same examinee: $X_h = 1$, and $X_g = 0$, for all $g < h$. Item-score combinations (0, 1) are called "errors" or "inversions." Permitted item-score patterns, (1, 0), (0, 0), and (1, 1), are known as Guttman patterns or "conformal" patterns. The general equation for group-based statistics is

$$G_i \equiv \frac{\sum_{g=1}^{r} w_g - \sum_{g=1}^{k} X_g w_g}{\sum_{g=1}^{r} w_g - \sum_{g=k-r+1}^{k} w_g} \,. \tag{3}$$

Person-fit statistics are often normed against a range of possible G_i values, given w_i. G_i is undefined for 0 and perfect score patterns. For these cases, $G_i = 0$: for these patterns, the perfect Guttman model holds. Thus, person-fit statistics that are based on group characteristics compare an individual's item-score pattern with the other item-score patterns in the sample.

Modified caution index. Harnisch & Linn's (1981) modified caution index C_i^* is a slight variation of Sato's (1975) caution index, C_i. C_i^* can be obtained from Equation 3 by setting $w_g = \pi_g$. C_i also is obtained by setting $w_g = \pi_g$ and multiplying $\sum_{g=k-r+1}^{k} w_g$ by r and the other terms by k. Both statistics weigh item scores with the proportion-correct score normed against the Guttman model. For example, $C_i^* = 0$ when an examinee with $X = r$ answers the r easiest items correctly and the $k - r$ most difficult items incorrectly. Thus, the examinee's item scores are in agreement with the Guttman model. Also, $C_i^* = 1$ when the item-score pattern equals the reversed Guttman pattern, indicating maximum misfit. The lower bound of C_i also equals 0 when an item-score pattern is in agreement with the Guttman model (Sato, 1975). However, C_i does not have a fixed upper bound, making its values more difficult to interpret than those of C_i^*.

Coefficients similar to C_i^* have been discussed (Cliff, 1983; Donlon & Fischer, 1968; Tatsuoka & Tatsuoka, 1982, 1983; van der Flier, 1977). Donlon and Fischer proposed the personal point-biserial correlation (r_{pbis}) as a person-fit statistic; r_{pbis} is the correlation across all items between an examinee's binary item scores and the vector containing the sample frequencies of the item scores (including the examinee's score). They also proposed the personal biserial correlation (r_{bis}), which

is the personal point-biserial correlation under the assumption of a continuous normally distributed variable underlying the binary item responses. Van der Flier defined his *U1* statistic as the number of Guttman errors normed against the maximum number of Guttman errors given $X = r$; this maximum equals $r(k - r)$.

Norm conformity and consistency indices. Tatsuoka & Tatsuoka (1983) discussed the norm conformity index (NCI_i),

$$NCI_i \equiv 1 - \frac{2\sum_{g=1}^{k-1}\sum_{h=g+1}^{k} X_g(1 - X_h)}{r(k - r)} . \tag{4}$$

The numerator contains the number of Guttman conformal (1, 0) item-score pairs multiplied by 2. In a reversed Guttman item-score vector, the number of conformal (1, 0) pairs equals 0, so $NCI_i = 1$. In a Guttman item-score vector, the number of conformal (1, 0) item-score pairs is $r(k - r)$, so $NCI_i = -1$. NCI_i is perfectly related to *U1*: $NCI_i = 1 - 2U1$.

Tatsuoka & Tatsuoka (1983) also discussed the individual consistency index (ICI_i). ICI_i is equivalent to NCI_i and is determined for subgroups of items that require the same cognitive solution strategy. Whereas NCI_i evaluates the consistency of an item-score pattern with the other score patterns in a group, ICI_i evaluates the consistency of an item-score pattern with an a priori defined item-score pattern based on the application of a particular cognitive skill.

Agreement, disagreement, and dependability indices. Kane & Brennan (1980) discussed agreement, disagreement, and dependability indices that can be used as group-based person-fit statistics. The agreement index is

$$A_i = \sum_{g=1}^{k} X_g \pi_g . \tag{5}$$

Let $A_i(\max)$ be the maximum value of A_i given the NC score r. $A_i(\max)$ is obtained if, given r, the item-score pattern is a Guttman pattern:

$$A_i(\max) = \sum_{g=1}^{r} \pi_g . \tag{6}$$

The disagreement index is

$$D_i = A_i(\max) - A_i . \tag{7}$$

The dependability index is

$$E_i = \frac{A_i}{A_i(\max)} . \tag{8}$$

Note that D_i equals the numerator of C_i^* (Equation 3, with $w_g = \pi_g$).

H_i^T statistic. Sijtsma (1986; Sijtsma & Meijer, 1992) proposed the person-fit statistic H_i^T. Let β_i be the expected proportion of items to which examinee i gives the correct response across locally independent repeated measurements, for a fixed set of k items. Let β_{ij} be the expected proportion of items to which examinees i and j respond correctly. Then $\sigma_{ij} = \beta_{ij} - \beta_i\beta_j$ is the covariance

between the scores of examinees i and j. If examinee indices $i < j$ imply $\beta_i \leq \beta_j$, the maximum covariance between the two examinees is obtained when $\beta_{ij} = \beta_i$; therefore,

$$\sigma_{ij}^{\max} = \beta_i(1 - \beta_j) \; . \tag{9}$$

For a single examinee in relation to $n - 1$ examinees,

$$H_i^T = \frac{\sum_{j \neq i} \sigma_{ij}}{\sum_{j \neq i} \sigma_{ij}^{\max}} \; . \tag{10}$$

The maximum value of H_i^T is 1 when each of the covariances between the item-score patterns of examinees i and j [for all i and $j (i \neq j)$] attains its maximum value. $H_i^T = 0$ when the average covariance (numerator) is 0; $H_i^T < 0$ if the average covariance is negative.

H_i^T is not normed against the Guttman pattern. Sijtsma (1986) showed that $H_i^T = 1$ is not necessary to obtain the perfect item-score pattern. Therefore, this statistic cannot be written in the form given in Equation 3.

U3 statistic. A group-based statistic with a known theoretical sampling distribution is van der Flier's (1980, 1982) *U3* statistic, which can be obtained from Equation 3 when

$$w_g = \ln\left(\frac{\pi_g}{1 - \pi_g}\right) \; . \tag{11}$$

To correct for dependence on an NC score, *U3* is standardized given $X = r$. This standardized statistic is

$$ZU3 = \frac{U3 - E(U3)}{[Var(U3)]^{1/2}} \; , \tag{12}$$

where $E(U3)$ and $Var(U3)$ are the expectation and the variance of *U3*, respectively. Van der Flier (1980, 1982) showed that, for long tests, *ZU3* is asymptotically standard normally distributed. When $X = r$, all terms in Equation 3 are constant, except $\sum_{g=1}^{k} X_g w_g$. Van der Flier (1982) derived expressions for $E(\sum_{g=1}^{k} X_g w_g)$ and $Var(\sum_{g=1}^{k} X_g w_g)$. The logit transformation of π_g enables the derivation of the normal distribution (see van der Flier, 1980, pp. 62–67).

Research on Group-Based Statistics

Studies comparing statistics. Harnisch & Linn (1981) used empirical data from reading and math tests to examine the correlations among C_i, C_i^*, r_{pbis}, r_{bis}, A_i , D_i, E_i, and NCI_i, and between these statistics and NC. Harnisch and Linn found that, for both tests, the correlations among all statistics except A_i were from .65 to .90. A_i correlated approximately .40 with each of the others. Most statistics correlated approximately .50 with NC on both tests. However, C_i^* correlated .20 (lowest) with NC, and A_i correlated .99. Harnisch and Linn then compared the average C_i^* scores across students for groups from different schools. They found significant inter-school differences that they attributed to differences in instruction and curriculum.

Rudner's (1983) simulation study compared the group-based statistics r_{pbis}, r_{bis}, NCI_i, and C_i with several IRT-based person-fit statistics: U, W, and l_0 (discussed below). High correlations (.61–.99) were found among the group-based statistics. Two cases were distinguished for investigating the effectiveness of the statistics in detecting misfitting item-score patterns. In one case, for a

minority of examinees, several correct responses were selected randomly and changed to incorrect responses, producing spuriously low NC scores. In a second case, incorrect responses were changed to correct responses, producing high NC scores. To identify whether the person-fit statistics could identify the altered item-score patterns, Rudner analyzed whether the spuriously high or low scores were correctly classified as misfitting by the statistics. Critical values were selected from a sample of 2,000 simulees. For each statistic, the critical value was determined by: (1) ordering the item-score patterns from decreasing to increasing misfit and (2) taking the value below which .05 of the most extreme values (indicating misfit) fell. Generally, the effectiveness of misfit detection increased with the number of altered items. For example, when 11% of responses were changed from incorrect to correct, NCI_i produced a detection rate of .10; when 33% of responses were similarly changed, NCI_i produced a detection rate of .20. Also, for tests consisting of 45 items, r_{bis} performed better than NCI_i and C_i. For longer tests (80 items), the IRT-based statistic U performed best.

Studies of a single person-fit statistic. Miller (1986) used C_i aggregated to the school class level to identify classes having a poor match between the content of a math test and instructional coverage. Differences in time spent on a particular subject in which students were to be tested resulted in different types of item mean values for C_i. A low C_i was found for classes in which test topics were emphasized, and a high C_i was found for classes in which other topics were emphasized.

Tatsuoka & Tatsuoka (1983) used NCI_i to detect deviant item-score patterns in an arithmetic test. They compared two groups of examinees: (1) students at a beginning level who made many different kinds of errors and (2) students near a mastery level who only made "sophisticated" errors. Item difficulties were different for the two groups, due to the different levels of expertise. Examinees making only sophisticated errors, but still included in the beginning level group, were classified as misfitting. When these same examinees were included in the mastery level group, they were classified as not misfitting. Thus, NCI_i obtained a relatively high value (indicating misfit) when an examinee's item scores deviated from the majority of item scores in the group. In the same study, ICI_i was used to identify examinees with inconsistent item-score patterns on items that required similar cognitive skills.

Jaeger (1988) used C_i^* to identify judges whose patterns of item judgment were misfitting in a standard-setting procedure (i.e., a procedure for establishing a decision rule for assigning candidates to pass/fail conditions). C_i^* ranged from .05–.62, with a mean of .32, and correlated .16 with the NC score on a reading test and .44 with the NC score on a mathematics test. Excluding judges with extreme C_i^* values had no effect on the recommended test standard.

Van der Flier (1982) simulated item-score patterns on the basis of item difficulties from two different populations (Populations I and II). *ZU3* scores were determined on the basis of π_g values in Population I or II. Item-score patterns were allocated to Population I or II on the basis of their *ZU3* scores and the significance probabilities in their corresponding populations. The exact decision rule that formed the basis for assigning a pattern to a population was unclear. Van der Flier found that approximately 70% of the patterns were allocated to the correct population, and the percentage of correct allocations was not related to the NC score.

Van der Flier (1982) then investigated the use of *ZU3* in a cross-cultural setting. Kenyan and Tanzanian examinees were compared on a Kiswahili verbal reasoning test. Kenyan examinees were known to have less knowledge of Kiswahili than Tanzanian examinees. Van der Flier hypothesized that: (1) for examinees with low *ZU3* scores (indicating misfit), the test scores would underestimate reasoning ability; and (2) for groups of examinees with equal test scores, a more deviant group would obtain better results on a criterion variable. It was found that Kenyans with large positive *ZU3* scores on verbal reasoning tests had better examination results (the criterion) than expected

based on their verbal reasoning test scores (the predictor). The additional information provided by the person-fit scores in predicting examination results, however, was rather modest.

Studies of detection rates. Meijer (1994), using simulated data, found that the detection rates of *U1* and *U3* were comparable. Also using simulated data, Meijer, Molenaar, & Sijtsma (1994) investigated the influence of test length, misfitting response types, and item discrimination on the detection rate of *U3*. They found that a priori defined misfitting item-score patterns were easier to detect in longer tests with higher item discriminations. Moreover, the type of misfitting behavior had a strong influence on the detection rate of *U3*. For example, misfitting item-score patterns were simulated by changing scores of 0 to 1 on the most difficult items (simulating cheating) or by assigning a probability of .25 to each item with a score of 1 (simulating guessing). Cheaters were easier to detect than guessers.

Meijer (1996) used simulated data to investigate the influence of the amount and type of misfitting patterns in a calibration sample on the detection rate of *ZU3*. As the number of misfitting simulees increased, the estimates of π_g were biased and the detection rate of *ZU3* decreased. Test length and type of misfit also influenced the detection rate. Re-estimating the proportion-correct score after removing misfitting patterns from the data using an iterative procedure also was investigated. This procedure was used until no further improvement in the detection rate was found. Results suggested that this method can be used to improve the detection rate of *ZU3* when there are misfitting persons.

Evaluation of Group-Based Statistics

In group-based person-fit statistics, a score pattern is classified as misfitting when items with proportion-correct near 0 are answered correctly, and items with proportion-correct near 1 are answered incorrectly. With the exception of *ZU3*, critical values for classifying item-response patterns as misfitting are usually chosen based on the characteristics of the data. For example, Harnisch (1983) suggested that a value higher than .6 indicated misfit for C_i, whereas Harnisch & Linn (1981) labelled item-score patterns with $C_i^* > .3$ as misfitting. These critical values, however, were based only on one or two empirical datasets.

Harnisch & Linn (1981) and Rudner (1983) used the following criteria to select useful person-fit statistics: (1) low correlation with the NC score and (2) detection rate. Harnisch and Linn concluded that, of the statistics considered in their study, C_i^* was related least to the NC score and was the most suitable statistic for detecting misfitting item-score patterns. A complete comparison of the correlations between person-fit statistics and test scores, and of the detection rates, seems impossible, however—the studies are incomplete, and the characteristics of the datasets are unclear.

Group-based statistics might be sensitive to misfitting item-score patterns, but their null distributions are unknown (with the exception of *ZU3*). As a result, significance probabilities cannot determine whether a score pattern is unlikely, given a nominal Type I error rate.

Let t be the observed value of a person-fit statistic T. Then, the significance probability, p^* (probability of exceedance), can be defined as the probability under the sampling distribution that the value of the test statistic is smaller or larger than the observed value [$p^* = P(T \leq t)$ and $p^* = P(T \geq t)$, respectively], depending on whether low or high values of the statistic indicate misfit. Although this might not be a serious problem when a person-fit statistic is used as a descriptive measure, the distribution of values for most group-based statistics is dependent on the test score (e.g., Drasgow, Levine, & McLaughlin, 1987). This dependence implies that, when a single critical value is used across test scores, the probability of classifying a score pattern as misfitting is a function of the test score, which is undesirable.

Group-based statistics such as *NCI* are similar to rank-order correlation measures [e.g., Kendall's (1970) τ], with two important differences. First, the values of Kendall's τ cannot be compared

across test scores. Second, the distributional properties of Kendall's τ are not easily applicable to those of person-fit statistics. In Cliff's (1996, pp. 66–88) overview of inferences based on ordinal correlation, the null hypotheses are different from those of interest here. For example, the null hypothesis that $\tau = 0$ for two variables can be tested against the alternative that there is positive or negative association. This null hypothesis is not very useful for person-fit research, because then it must be determined whether the item scores follow the Guttman model (implying perfect association; the alternative would be no perfect association, but often positive covariance).

IRT-Based Person-Fit Measures

Statistics

Prerequisites. In IRT, the probability of correctly answering item g ($g = 1, 2, \ldots, k$) is a function of θ and item characteristics (e.g., δ; Embretson & Reise, 2000; Hambleton & Swaminathan, 1985; van der Linden & Hambleton, 1997). This conditional probability $P_g(\theta)$ is the item response function. A vector with item-score random variables is $\mathbf{X} = (X_1, X_2, \ldots, X_k)$, and a realization is $\mathbf{x} = (x_1, x_2, \ldots, x_k)$. IRT often assumes that the item scores are locally independent,

$$P(\mathbf{X} = \mathbf{x}|\theta) = \prod_{g=1}^{k} P_g(\theta)^{x_g}[1 - P_g(\theta)]^{1-x_g} . \tag{13}$$

For any cumulative probability distribution $F(\theta)$, θ can be integrated out, resulting in

$$P(\mathbf{X} = \mathbf{x}) = \int_{\theta} \prod_{g=1}^{k} P_g(\theta)^{x_g}[1 - P_g(\theta)]^{1-x_g} dF(\theta) . \tag{14}$$

For testable restrictions on the distribution of $\mathbf{X}$, specific choices for $P_g(\theta)$, $F(\theta)$, or both must be made. Although $F(\theta)$ sometimes is selected to be normal, $P_g(\theta)$ often is specified using the one-, two-, or three-parameter logistic model (1PLM, 2PLM, and 3PLM, respectively). For the 3PLM,

$$P_g(\theta) = \gamma_g + \frac{(1 - \gamma_g)\exp[\alpha_g(\theta - \delta_g)]}{1 + \exp[\alpha_g(\theta - \delta_g)]} , \tag{15}$$

where

γ_g is the lower asymptote (γ_g is the probability of a 1 score for persons with low θs—that is, $\theta \to -\infty$),
α_g is the slope (item discrimination) parameter, and
δ_g is the item location parameter.

The 2PLM can be obtained by fixing $\gamma_g = 0$ for all items, and the 1PLM or Rasch (1960/80) model by additionally fixing $\alpha_g = 1$ for all items.

A major advantage of IRT models is that a model's goodness of fit to empirical data can be investigated. Compared to group-based person-fit statistics, this provides the opportunity of evaluating the item-score pattern fit to an IRT model.

Following Snijders (in press), a general form in which most person-fit statistics can be expressed is

$$\sum_{g=1}^{k} X_g w_g(\theta) - w_0(\theta) , \tag{16}$$

where $w_g(\theta)$ and $w_0(\theta)$ are suitable functions for weighting the item scores and adapting person-fit scale scores, respectively.

So that the expectation of a person-fit statistic is 0, many person-fit statistics are expressed in the centered form

$$V = \sum_{g=1}^{k} [X_g - P_g(\theta)] w_g(\theta) \ . \tag{17}$$

Note that, as a result of binary scoring, $X_g^2 = X_g$. Thus, for a suitable function $v_g(\theta)$,

$$V^* = \sum_{g=1}^{k} [X_g - P_g(\theta)]^2 v_g(\theta) \tag{18}$$

can be re-expressed as Equation 16.

Residual-based statistics. Wright & Stone (1979) and Wright & Masters (1982, pp. 108–111) proposed two mean-squared residual-based statistics, U and W. U is based on squared standardized residuals. The weight

$$v_g(\theta) = \frac{1}{k P_g(\theta)[1 - P_g(\theta)]} \tag{19}$$

results in

$$U = \sum_{g=1}^{k} \frac{[X_g - P_g(\theta)]^2}{k P_g(\theta)[1 - P_g(\theta)]} \ . \tag{20}$$

The denominator of these equations contains the conditional variances of the individual item scores: $Var(X_g|\theta) = P_g(\theta)[1 - P_g(\theta)]$. U can be interpreted as the mean of the squared standardized residuals based on k items. Further,

$$W = \frac{\sum_{g=1}^{k} [X_g - P_g(\theta)]^2}{\sum_{g=1}^{k} P_g(\theta)[1 - P_g(\theta)]} \ . \tag{21}$$

Wright & Stone (1979) assumed that W is less sensitive than U to an unexpected response to an item with a difficulty distant from an examinee's θ. According to Wright and Stone and Wright & Masters (1982),

$$ZU = [\ln U + U + 1](df/8)^{-1} \ , \tag{22}$$

with $k - 1$ degrees of freedom (df). Similarly,

$$ZW = 3(W^{1/3} - 1)/q + (q/3) \ , \tag{23}$$

where q is the variance of W. Wright and Stone and Wright and Masters claimed that these transformations are asymptotically standard normally distributed. The appropriateness of ZU and

ZW for approximating the normal distribution can be questioned, however, as will be discussed below.

Smith (1985) proposed two related statistics. He assumed that a test can be divided into S nonoverlapping subsets of items, $A_s (s = 1, 2, \ldots, S)$. He then defined an unweighted between-sets fit statistic as

$$UB = \frac{1}{S-1} \sum_{s=1}^{S} \frac{\left\{ \sum_{g \in A_s} [X_g - P_g(\theta)] \right\}^2}{\sum_{g \in A_s} P_g(\theta)[1 - P_g(\theta)]} . \tag{24}$$

Let m_s be the number of items in subset A_s; then, the unweighted within-sets fit statistic is

$$UW = \frac{1}{m_s} \sum_{g \in A_s} \frac{[X_g - P_g(\theta)]^2}{k P_g(\theta)[1 - P_g(\theta)]} . \tag{25}$$

Smith (1985, 1986) used critical values obtained from a simulation study for classifying examinees as model-fitting or misfitting. For the Rasch model, Kogut (1988) showed that (1) the joint distribution of subtest residuals,

$$\frac{\sum_{g \in A_s} [X_g - P_g(\theta)]}{\sum_{g \in A_s} P_g(\theta)[1 - P_g(\theta)]} , \tag{26}$$

is asymptotically multivariate normal; and (2) UB has an asymptotically χ^2 distribution with S df when θ is used, and $S - 1$ df when the maximum likelihood estimate (MLE) $\hat{\theta}$ is used. Empirical distributions were simulated to investigate whether the asymptotic distributions held reasonably well for tests of realistic length (40 items; Kogut, 1988). The empirical distributions were accurate enough to approximate the asymptotic distributions. UW can investigate whether a priori specified subsets of items fit the IRT model.

Likelihood-based statistics. The log-likelihood function

$$l_0 = \sum_{g=1}^{k} \{X_g \ln P_g(\theta) + (1 - X_g) \ln[1 - P_g(\theta)]\} , \tag{27}$$

first used by Levine & Rubin (1979) to assess person fit, has been further developed and applied (e.g., Drasgow, Levine, & McLaughlin, 1991; Drasgow, Levine, & Williams, 1985; Levine & Drasgow, 1982, 1983). l_0 is determined as the logarithm of the likelihood function evaluated at the MLE of θ. Two problems occur when using l_0 as a fit statistic.

1. l_0 is not standardized, implying that the classification of an item-score pattern as model-fitting or misfitting depends on θ.
2. For classifying an item-score pattern as misfitting, a distribution of l_0 under the null hypothesis of fitting item scores is needed. This null distribution is unknown for l_0.

To overcome these problems, Drasgow et al. (1985) proposed a standardized version of l_0 that is less confounded with θ and purported to be asymptotically standard normally distributed. This standardized version of l_0 is

$$l_z = \frac{l_0 - E(l_0)}{[Var(l_0)]^{1/2}} , \tag{28}$$

where $E(l_0)$ and $Var(l_0)$ are the expectation and variance of l_0, respectively:

$$E(l_0) = \sum_{g=1}^{k} \{P_g(\theta) \ln[P_g(\theta)] + [1 - P_g(\theta)] \ln[1 - P_g(\theta)]\} , \tag{29}$$

and

$$Var(l_0) = \sum_{g=1}^{k} P_g(\theta)[1 - P_g(\theta)] \left[\ln \frac{P_g(\theta)}{1 - P_g(\theta)}\right]^2 . \tag{30}$$

Molenaar & Hoijtink (1990, 1996) argued that l_z is standard normally distributed only when true θ values are used. A problem arises in practice when θ is replaced by the MLE, $\hat{\theta}$. Using an estimate instead of true θ has an effect on the distribution of a person-fit statistic (Molenaar & Hoijtink, 1990; Nering, 1995, 1997; Reise, 1995). When $\hat{\theta}$ is used, the variance of l_z is smaller than expected under the standard normal distribution using the true θ, particularly for tests of moderate length (e.g., 50 items or fewer). The empirical Type I error was found to be smaller than the nominal Type I error. This effect could not be reduced using Warm's θ estimator, which corrects for overestimated positive values and underestimated negative values of θ (van Krimpen-Stoop & Meijer, 1999).

For the Rasch model, Molenaar & Hoijtink (1990, p. 96) showed that l_0 can be written as the sum of two terms,

$$l_0 = d_0 + M , \tag{31}$$

with

$$d_0 = -\sum_{g=1}^{k} \ln[(1 + \exp(\theta - \delta_g)] + r\theta , \tag{32}$$

and

$$M = -\sum_{g=1}^{k} \delta_g X_g . \tag{33}$$

Given $\sum_{g=1}^{k} X_g = r$ (that is, given $\hat{\theta}$, which in the Rasch model depends only on the sufficient statistic r), d_0 is independent of the item-score pattern $\mathbf{X}$, and M is dependent on it. l_0 and M have the same ordering in $\mathbf{X}$.

Because of its simplicity, Molenaar & Hoijtink (1990) used M rather than l_0 as a person-fit statistic. They proposed three approximations to the distribution of M: (1) complete enumeration; (2) monte carlo simulation; and (3) a χ^2 distribution, in which the mean, standard deviation (SD), and skewness of M are taken into account. (See Molenaar & Hoijtink, 1990, for when these approaches should be used; Liou & Chang, 1992, for a network algorithm that enumerates all possible response patterns to construct exact tail probabilities for l_0; and Bedrick, 1997, for alternative methods to approximate the first two moments of M.)

Group-based statistics such as G (Equation 3) can be used to detect misfitting item-score patterns under an IRT model. However, they might classify different item-score patterns as misfitting in comparison to fit statistics based on IRT parameters. Molenaar & Hoijtink (1990) argued that calculating a distribution for statistics such as G is equally laborious as for M, which was designed especially for detecting misfit under the Rasch model. Molenaar and Hoijtink examined possible discrepancies between M and other statistics. They found that many model-fitting response patterns identified using M were identified as misfitting using (a weighted version of) the number of Guttman inversions. Thus, G is not recommended in the context of the Rasch model.

Drasgow et al. (1991) proposed a generalization of l_z for tests consisting of S unidimensional subtests. This statistic has a form similar to l_z, but the expectation and variance are taken over S subtests,

$$l_{zm} = \frac{\sum_{s=1}^{S} \left\{ l_0^{(s)} - E[l_0^{(s)}] \right\}}{\sum_{s=1}^{S} \left\{ Var[l_0^{(s)}] \right\}^{1/2}} . \tag{34}$$

Although l_{zm} was effective in detecting misfitting item-score patterns, detection rates were approximately equal to those for long, unidimensional tests with a number of items equaling the total number of items in the S subtests. In practical testing situations, l_{zm} has the same problems as l_z: $\hat{\theta}$ is used, resulting in inappropriate approximations to probabilities of exceedance. Using l_z with the 3PLM, Nering (1995) found that the empirical Type I error in general was lower than the nominal Type I error.

Snijders (1998) derived the asymptotic sampling distribution for a group of person-fit statistics using $\hat{\theta}$ instead of θ and taking the form of Equation 17. Snijders showed that $l_0 - E(l_0)$ can be written in the form of Equation 17 when

$$w_g = \frac{P_g(\theta)}{1 - P_g(\theta)} . \tag{35}$$

Snijders (in press) derived expressions for the first two moments: $E[V(\hat{\theta})]$ and $Var[V(\hat{\theta})]$. A simulation study then was performed for relatively small tests (8 and 15 items) using the 2PLM and $\hat{\theta}$. The approximation was satisfactory for $\alpha = .05$ and $\alpha = .10$, but the empirical Type I error was higher than the nominal Type I error for smaller values of α.

Drasgow et al. (1987) proposed two fit statistics that are sensitive to the flatness of the likelihood function. They indicated that, when there is no single value of θ providing a good fit for an item-score pattern, the likelihood function will be relatively flat. The first statistic is a normalized jackknife variance estimate (JK). If $\hat{\theta}$ is the usual 3PLM MLE of θ based on all k items and $\hat{\theta}^*_{(g)}$ is the 3PLM MLE of θ based on the $k - 1$ items remaining when item g is excluded, then

$$\hat{\theta}^*_g \equiv k\hat{\theta} - (k-1)\hat{\theta}^*_{(g)} , \quad g = 1, 2, \ldots , k . \tag{36}$$

The jackknife estimate of θ is

$$\hat{\theta}^* = 1/k \sum_{g=1}^{k} \hat{\theta}^*_g , \tag{37}$$

with

$$Var(\hat{\theta}^*) = \frac{\sum_{g=1}^{k}(\hat{\theta}_g^*)^2 - 1/k\left(\sum_{g=1}^{k}\hat{\theta}_g^*\right)^2}{k(k-1)} . \tag{38}$$

Because there is more Fisher information about θ within certain ranges, $Var(\hat{\theta}^*)$ depends on θ. It therefore is weighted by the Fisher information, $I(\hat{\theta})$, the reciprocal of which is the asymptotic variance of $\hat{\theta}$. This results in

$$JK = Var(\hat{\theta}^*)I(\hat{\theta}) . \tag{39}$$

When an item-score pattern does not fit the model, the likelihood function is relatively flat and the variance estimate is higher than for a fitting item-score pattern. Drasgow et al.'s (1987) second person-fit statistic is the ratio of observed and expected information,

$$O/E = \frac{\left.\frac{\partial^2 l_0}{\partial\theta^2}\right|_{\theta=\hat{\theta}}}{I(\hat{\theta})} . \tag{40}$$

When the likelihood l_0 (see Equation 27) is flatter for misfitting responses than for model-fitting responses, the observed information is expected to be smaller than the expected information.

Statistics based on the caution index. Tatsuoka & Linn (1983) developed several person-fit statistics similar to the caution index C_i (Harnisch & Linn, 1981) that use IRT modeling. The caution index can be written as

$$C_i = 1 - \frac{Cov(\mathbf{X}_i, \mathbf{n})}{Cov(\mathbf{X}_i^*, \mathbf{n})} , \tag{41}$$

where

$\mathbf{X}_i$ is a vector of item scores of examinee i,
$\mathbf{X}_i^*$ is the theoretical Guttman vector, and
$\mathbf{n}$ is a vector with the item NC scores across examinees.

By norming against the covariance between the probability of a correct response under an IRT model and $\mathbf{n}$, *ECI1* was obtained as

$$ECI1 = 1 - \frac{Cov(\mathbf{X}_i, \mathbf{n})}{Cov[\mathbf{P}(\theta), \mathbf{n}]} , \tag{42}$$

where $\mathbf{P}(\theta)$ is a vector defined for each θ with conditional probabilities $P_g(\theta)$ across items.

ECI2 and *ECI3* were obtained by computing the covariance (correlation) between an item-score vector and the vector with the mean probability of correctly answering an item across n examinees,

$$ECI2 = 1 - \frac{Cov[\mathbf{X}_i, \mathbf{G}]}{Cov[\mathbf{P}(\theta), \mathbf{G}]} , \tag{43}$$

and

$$ECI3 = 1 - \frac{Corr[\mathbf{X}_i, \mathbf{G}]}{Corr[\mathbf{P}(\theta), \mathbf{G}]} , \tag{44}$$

where $G = (G_1, G_2, \ldots, G_k)$ with elements $G_g = 1/n \sum_{i=1}^{n} P_g(\theta)$.

ECI4, *ECI5*, and *ECI6* were obtained by computing the covariance or the correlation between the response vector $\mathbf{X}_i$ and $\mathbf{P}(\theta)$, resulting in

$$ECI4 = 1 - \frac{Cov[\mathbf{X}_i, \mathbf{P}(\theta)]}{Cov[\mathbf{G}, \mathbf{P}(\theta)]} , \tag{45}$$

$$ECI5 = 1 - \frac{Corr[\mathbf{X}_i, \mathbf{P}(\theta)]}{Corr[\mathbf{G}, \mathbf{P}(\theta)]} , \tag{46}$$

and

$$ECI6 = 1 - \frac{Cov[\mathbf{X}_i, \mathbf{P}(\theta)]}{Var[\mathbf{P}(\theta)]} . \tag{47}$$

Although *ECI2* and *ECI3* compare an individual item-score pattern with the *mean* probability across persons—thus comparing an individual item-score pattern with group characteristics—*ECI4*, *ECI5*, and *ECI6* compare an individual item-score pattern with the expected probability on the basis of a model. *ECI4* is normed against the mean probability across items; *ECI6* is normed against the variance of $\mathbf{P}(\theta)$. *ECI3* and *ECI5* are similar to *ECI2* and *ECI4*, except that in *ECI3* and *ECI5*, correlations are used instead of covariances. Tatsuoka (1984) derived the expectations and variances of *ECI1*, *ECI2*, *ECI4*, and *ECI5* to obtain standardized versions of these indices (subtracting the expected values and dividing by the SDs). These standardized indices were denoted $ECI1_z$, $ECI2_z$, $ECI4_z$, and $ECI5_z$.

Although likelihood and *ECI* statistics are based on different approaches to person fit (e.g., Harnisch & Tatsuoka, 1983; Kogut, 1986; Nering, 1997), both approaches are of the form of Equation 17. For example, the centered form of *ECI4*, $ECI4 - E(ECI4)$, can be obtained when

$$w_g = P_g(\theta) - \bar{P}(\theta) , \tag{48}$$

where $\bar{P}(\theta) = 1/k \sum_{g=1}^{k} P_g(\theta)$. The centered form of *ECI2* can be obtained when

$$w_g = G_g - \overline{G} , \tag{49}$$

where

$$\overline{G} = 1/k \sum_{g=1}^{k} G_g . \tag{50}$$

Optimal Person-Fit Statistics

Levine & Drasgow (1988; Drasgow & Levine, 1986; Drasgow, Levine, & Zickar, 1996) proposed a statistically optimal method for the identification of misfitting item-score patterns: no other method can achieve a higher rate of detection at the same Type I error rate. A likelihood ratio statistic was proposed that provides the most powerful test for the null hypothesis that an item-score pattern is model-fitting (versus misfitting). A model for model-fitting behavior (e.g., 1-, 2-, or 3PLM) and a model for a particular type of misfitting behavior (e.g., a model with violations of local independence) are specified in advance. The likelihood ratio statistic used is

$$\lambda(\mathbf{X}) = \frac{P(\mathbf{X} = \mathbf{x})_{misfitting}}{P(\mathbf{X} = \mathbf{x})_{fitting}} . \tag{51}$$

Response patterns are classified as misfitting if they have (1) the largest $\lambda(\mathbf{X})$ and (2) likelihoods under the model describing model-fitting response behavior that sum to the α level.

Klauer (1991, 1995) investigated misfitting item-score patterns by testing a null model of model-fitting response behavior (Rasch model) against an alternative model of misfitting response behavior. Writing the Rasch model as a member of the exponential family,

$$P(\mathbf{X} = \mathbf{x}|\theta) = \mu(\theta)h(\mathbf{x}) \exp[\theta R(\mathbf{x})] , \tag{52}$$

where

$$\mu(\theta) = \prod_{g=1}^{k} [1 + \exp(\theta - \delta_g)]^{-1} , \tag{53}$$

and

$$h(\mathbf{x}) = \exp(-\sum x_g \delta_g) , \tag{54}$$

where $R(\mathbf{x})$ is an NC score. Klauer (1995) modeled misfitting response behavior using the two-parameter exponential family with an extra person parameter, η, as

$$P(\mathbf{X} = \mathbf{x}|\theta, \eta) = \mu(\theta, \eta)h(\mathbf{x}) \exp[\eta T(\mathbf{x}) + \theta R(\mathbf{x})] , \tag{55}$$

where $T(\mathbf{x})$ depends on the particular alternative model considered. Using the exponential family of models, a uniformly most powerful test (Lindgren, 1993, p. 350) can be used for testing $H_0{:}\eta = \eta_0$ against $H_1{:}\eta \neq \eta_0$. Let a test be subdivided into two subtests, A_1 and A_2. Then, as an example of η, $\eta = \theta_1 - \theta_2$ was considered, where θ_1 is an examinee's θ on subtest A_1, and θ_2 is an examinee's θ on subtest A_2. Under the Rasch model, θ is expected to be invariant across subtests; therefore, $H_0{:}\eta = 0$ can be tested against $H_1{:}\eta \neq 0$. For this type of misfitting behavior, $T(\mathbf{x})$ is the NC score on either of the subtests.

Klauer (1995) also tested the H_0 of equal item discrimination parameters for all examinees against person-specific item discrimination and the H_0 of local independence against violations of local independence. Klauer found that the power of these tests depended on the type and severity of the violations. Violations against noninvariant θ ($H_0{:}\eta = 0$) were found to be the most difficult to detect. Liou (1993) discussed refinements using these types of tests.

Levine & Drasgow (1988) and Klauer (1991, 1995) specified model violations in advance, and tests were proposed to investigate these violations. In most person-fit studies, the alternative hypothesis is simply the logical reverse of the null hypothesis. An obvious problem is which alternative models to specify. A possibility is to specify a number of plausible alternative models and then successively test model-conformed item-score patterns against those alternatives. Another option is to investigate which model violations are most detrimental to the use of the chosen test, and then test against the most serious violations (Klauer, 1995).

The Person Response Function

Trabin & Weiss (1983; see also Weiss, 1973) proposed using the person response function (PRF) to identify misfitting item-score patterns. At a fixed θ value, the PRF specifies the probability of a correct response as a function of the item location, δ (and other item parameters). In IRT, the item response function for dichotomously scored items is assumed to be a nondecreasing function of θ, whereas the PRF is assumed to be a nonincreasing function of δ (Trabin & Weiss, 1983). To construct

an observed PRF, Trabin and Weiss ordered items from low to high $\hat{\delta}$ and then formed subtests by grouping items according to $\hat{\delta}$. For fixed $\hat{\theta}$, the observed PRF was constructed by determining the proportion of correct responses in each subtest. The expected PRF was constructed by estimating, according to the 3PLM, the mean probability of a correct response. A large difference between the expected and observed PRFs was interpreted as an indication of misfitting responses for that examinee.

Let k items be ordered by their δ, and let item rank numbers be assigned such that $\delta_1 < \delta_2 < \ldots < \delta_k$. Assume that S ordered subtests $A_s (s = 1, 2, \ldots, S)$ can be formed, each containing m items. Then, $A_1 = \{1, 2, \ldots, m\}$, $A_2 = \{m + 1, \ldots, 2m\}$, $\ldots$, $A_S = \{k - m + 1, \ldots, k\}$, and $S \times m = k$. To construct the expected PRF, an estimate of the expected proportion of correct responses under the 3PLM in each subtest is

$$m^{-1} \sum_{g \in A_s} P_g(\hat{\theta}) , \quad s = 1, 2, \ldots, S . \tag{56}$$

This expected proportion is compared with the observed proportion of correct responses, given by

$$m^{-1} \sum_{g \in A_s} X_g , \quad s = 1, 2, \ldots, S . \tag{57}$$

For a particular $\hat{\theta}$ within each subtest, the difference between observed and expected correct scores is taken and divided by the number of items in the subtest, yielding

$$D_s(\hat{\theta}) = m^{-1} \sum_{g \in A_s} \left[X_g - P_g(\hat{\theta}) \right] , \quad s = 1, 2, \ldots, S . \tag{58}$$

Ds then are added across subtests, yielding

$$D(\hat{\theta}) = \sum_{s=1}^{S} D_s(\hat{\theta}) . \tag{59}$$

$D(\hat{\theta})$ is a measure of an examinee's fit to the model. For example, when examinees copy answers to the most difficult items from another examinee, their scores on those subtests are likely to be substantially higher than predicted by the expected PRF. (For related ideas, see Lumsden, 1977, 1978.)

Klauer & Rettig (1990) expanded the methodology of Trabin & Weiss (1983) by proposing three standardized person-fit statistics that asymptotically follow a χ^2 distribution for long tests. One of their statistics is

$$\chi^2_{sc} = \sum_{s=1}^{S} \frac{V_s^2(\hat{\theta})}{I_s(\hat{\theta})} , \tag{60}$$

where $V_s(\hat{\theta})$ is (based on Equation 17)

$$V_s(\hat{\theta}) = \sum_{g \in A_s} \left[X_g - P_g(\hat{\theta}) \right] w_g(\hat{\theta}) , \tag{61}$$

where

$$w_g(\theta) = \frac{dP_g(\theta)/d\theta}{P_g(\theta)[1 - P_g(\theta)]} , \tag{62}$$

and $I_s(\hat{\theta})$ is the estimated Fisher information function. To determine whether θ is invariant across subtests, the null hypothesis H_0:$\theta_1 = \theta_2 = \ldots = \theta_S$ is tested. Under H_0, χ^2_{sc} has $df = S - 1$. Although similar to Trabin & Weiss's (1983) method, χ^2_{sc} is standardized and asymptotically χ^2 distributed.

Klauer & Rettig (1990) also proposed two related tests. The first was the Wald test, which compares an examinee's $\hat{\theta}$s from different subtests. The other was a likelihood ratio test. Through monte carlo research, Klauer and Rettig showed that χ^2_{sc} was distributed as χ^2 for tests of at least 80 items. For the Wald and likelihood ratio tests, the difference between the theoretical and empirical χ^2 distributions was too large to be of practical use.

Sijtsma & Meijer (in press) proposed a PRF approach to person-fit research in a nonparametric IRT context. They found that the nonparametric PRF approach can provide diagnostic information about the type of misfit; it can be used in addition to an overall person-fit statistic, which only identifies misfitting item-score patterns.

Person-Fit Research in Computerized Adaptive Testing

With the increasing use of computerized adaptive testing (CAT; Meijer & Nering, 1999; Weiss, 1982), person-fit statistics might be helpful for detecting item memorization or examinees who are familiar with some of the items (due to continuous test administration from the same item bank). In CAT, the distributional characteristics of existing person-fit statistics (e.g., l_z and *ECI4*) do not agree with the expected theoretical distributions (McLeod & Lewis, 1998, 1999; Nering, 1997; van Krimpen-Stoop & Meijer, 1999). This might be explained by the modest variability in item difficulties in a CAT, which results in reduced variability of the assumed null distribution of person-fit statistics. Consequently, empirical Type I errors are small compared to nominal Type I errors (van Krimpen-Stoop & Meijer, 1999). Person-fit statistics designed especially for CAT might be more powerful than "conventional" person-fit statistics.

McLeod & Lewis (1999) proposed the Z_c statistic for detecting item-score patterns that result from memorization. Before Z_c is calculated, the item bank is divided into three parts: easy, medium difficulty, and difficult items. Let $Easy[P_g(\theta) - X_g]$ denote the mean residual for the easy items, and $Diff[P_g(\theta) - X_g]$ the mean residual for the most difficult items in a CAT. Then,

$$Z_c = \frac{Easy[P_g(\theta) - X_g] - Diff[P_g(\theta) - X_g]}{\left\{\sum_{Easy}\{P_g(\theta)[1 - P_g(\theta)]\}/n^2_{Easy}\right\} + \left\{\sum_{Diff}\{P_g(\theta)[1 - P_g(\theta)]\}/n^2_{Diff}\right\}} . \quad (63)$$

Z_c is positive when an examinee incorrectly answers easy items and correctly answers difficult items. However, applying Z_c to an operational Graduate Record Examination Quantitative CAT bank with 14 memorized items resulted in low detection rates.

Bradlow, Weiss, & Cho (1998) and van Krimpen-Stoop & Meijer (2000) proposed person-fit statistics in which a model-fitting item-score pattern consists of an alternation of correct and incorrect responses, especially at the end of the test when $\hat{\theta}$ converges on θ. A string of consecutive correct or incorrect answers could indicate misfit. Sums of consecutive negative or positive residuals $[P_g(\theta) - X_g]$ can be investigated using a cumulative sum procedure (Page, 1954). For each item g in the test, a statistic T_g can be calculated that is a weighted version of $[P_g(\theta) - X_g]$. Then, the sum of these T_gs is

$$C^+_g = \max[0, T_g + C^+_{g-1}] , \quad (64)$$

$$C^-_g = \min[0, T_g + C^-_{g-1}] , \quad (65)$$

and

$$C_0^+ = C_0^- = 0\ , \tag{66}$$

where C^+ and C^- reflect the sum of consecutive positive and negative residuals, respectively. Let UB and LB be appropriate upper and lower bounds, respectively. Then, when $C^+ > UB$ or $C^- < LB$, the item-score pattern can be classified as not fitting the model; otherwise, the item-score pattern can be classified as fitting.

Research Using IRT-Based Person-Fit Statistics

Studies using simulated or empirical data and addressing the usefulness of IRT-based person-fit statistics have investigated:

1. The detection rate of fit statistics, and comparison of fit statistics with respect to several criteria (e.g., distributional characteristics, relation to test scores).
2. The influence of item, test, and person characteristics on the detection rate.
3. The applicability of person-fit statistics for detecting particular types of misfitting item-score patterns.
4. The relation between misfitting score patterns and the validity of test scores.

Although some studies can be categorized under more than one heading, they are discussed under the heading for which they appear to have made their largest contribution.

Detection Rate and Fit-Statistic Comparison

Levine & Rubin (1979) evaluated l_0 in a study simulating item-score vectors using item parameters estimated from the Scholastic Aptitude Test (Verbal). Spuriously high-scoring examinees were simulated by randomly sampling a fixed percentage of the item scores of model-fitting examinees (generated using the 3PLM) and changing these scores to correct. Spuriously low-scoring examinees similarly were simulated by rescoring randomly selected items as incorrect with a probability of .20. For each group, 4%, 10%, 20%, and 40% of the item responses were changed. Levine and Rubin found that the larger the group of misfitting item scores, the better l_0 could distinguish fitting from misfitting score patterns. They also found that spuriously high-scoring simulees were easier to detect than spuriously low-scoring simulees, because (as a result of the procedures they used for changing scores) more item scores were changed for those with spuriously high scores.

Drasgow (1982) compared the detection rates of l_0 using either the Rasch model or the 3PLM with data from the Graduate Record Examination. He also found higher detection rates for examinees with spuriously low manipulated item scores than for those with spuriously high manipulated item scores. Detection rates for this dataset were higher using the 3PLM than the Rasch model.

Harnisch & Tatsuoka (1983) used National Assessment of Educational Progress (NAEP) data on mathematics to investigate distributional characteristics and relationships among several *ECI* indices, U, W, l_0, and l_z. U was used under the 2PLM and 3PLM, and l_0 was used under the 3PLM and normal-ogive model (Hambleton & Swaminathan, 1985, pp. 35–36). Harnisch and Tatsuoka found that *ECI1*, *ECI2*, and *ECI4* had SDs of approximately 1 and means of approximately .20. U correlated lowest with the other statistics (approximately .10), and the other statistics correlated between .50 and .98 with each other. l_z and l_0 correlated highest with the NC score (.36 and .27, respectively). The strongest curvilinear relationships were between the NC score and l_0 and W.

Drasgow et al. (1987) used the 3PLM for comparing l_z, *ZU*, *ZW*, *C*, *JK*, *O/E*, $ECI2_z$, and $ECI4_z$ (1) to optimal statistics, (2) on their numerical values across θ, and (3) with regard to their detection

rates. Similar to Levine & Rubin (1979), detection rates were found by determining the proportions of misfitting item-score patterns correctly identified as misfitting. Drasgow et al. concluded that *ZU* and *C* were poorly standardized compared to the other indices. They also found that $ECI4_z$ was better standardized and had a higher detection rate than $ECI2_z$. *O/E* and *JK* were reasonably well standardized, but they were ineffective for detecting misfit. l_z, *ZW*, and $ECI4_z$ had high detection rates for spuriously high-scoring examinees with low θs and spuriously low-scoring examinees with high θs (e.g., the detection rate of $ECI4_z$ was .75 for low θs at a Type I error of .01 for 30 spuriously high-scoring examinees). However, these statistics were less sensitive to manipulated response patterns for θs near 0 (e.g., for $ECI4_z$, the detection rate decreased from .75 to .51 for low θs at a Type I error of .01). Optimal indices had detection rates ranging from 50–200% higher than other indices for average θs and item-score patterns with spuriously high or low test scores.

Rogers & Hattie (1987; see also Reise, 1990) investigated the detection rates of *ZU* and *ZW*. Transformations of both statistics were claimed (Wright & Stone, 1979) to be asymptotically standard normally distributed. Rogers and Hattie found the detection rates of *ZU* and *ZW* using theoretical critical values for: (1) guessing, (2) heterogeneity of the discrimination parameters, and (3) multidimensionality. They concluded that *ZW* was insensitive to heterogeneity and multidimensionality and sensitive to guessing. *ZU* was insensitive to all three types of misfit. Detection rates increased by no more than 2%, compared to detection rates for model-fitting examinees.

Noonan, Boss, & Gessaroli (1992) investigated the distributional characteristics and empirical critical values of l_z, $ECI4_z$, and *ZW* as functions of test length and IRT model (2PLM and 3PLM). They found that l_z and $ECI4_z$ had means and SDs that approximated the standard normal distribution. However, *ZW* had a mean over replications of approximately 1.00, but an SD between .144 and .232. $ECI4_z$ and *ZW* were positively skewed, and l_z was negatively skewed; the skewness of $ECI4_z$ was half the skewness of the other two statistics. For all three statistics, the critical values were affected by test length and IRT model, with *ZW* most affected. They concluded that $ECI4_z$ best approximated the normal distribution and was least affected by test length and IRT model. l_z and $ECI4_z$ were highly correlated (.95), whereas $ECI4_z$ and *ZW* had the lowest correlation (.58). However, true θ values were used, which makes generalizations to empirical distributions difficult.

Li & Olejnik (1997) compared the distributions of l_z, $ECI2_z$, $ECI4_z$, *ZU*, and *ZW* using the Rasch model. They found that (1) the statistics had low correlations with NC scores; (2) the statistics were positively skewed and deviated significantly from normality ($ECI4_z$ was better normalized than $ECI2_z$); (3) l_z performed at least as well as the other statistics in detecting misfitting behavior; (4) examinees with spuriously low and spuriously high NC scores were equally well detected when unidimensional data were used, whereas detection rates of spuriously low NC scores were lower than detection rates of spuriously high NC scores when a multidimensional test was used; and (5) person-fit statistics were not very powerful in identifying misfitting item-score patterns (l_z was most powerful and detected at most 67% of the misfitting item-score patterns). However, it was assumed that the true θ equaled the MLE $\hat{\theta}$. As discussed above, when using the Rasch model, it is better to condition on the NC score, which is independent of $\hat{\theta}$, and to use the *M* statistic. This was done by Kogut (1987), who used simulated Rasch model data to show that the detection rate of *M* for detecting misfitting item-score patterns was higher than l_z.

Trabin & Weiss (1983) applied the PRF approach to a 216-item vocabulary test administered to 151 undergraduate students. To investigate whether the responses were in agreement with the 3PLM, they used $D(\theta)$ to evaluate the discrepancy between the observed and expected PRFs for each student, and assumed that $D(\theta)$ was χ^2 distributed. Some students had significant χ^2s, but the cause of misfit could not be explained.

Nering & Meijer (1998) used simulated data for comparing the PRF approach with l_z. They found that the detection rate of l_z was higher than that of the PRF method in most cases. They suggested that the PRF approach and l_z can be used in a complementary way: misfitting item-score patterns can be detected using l_z, and differences between expected and observed PRFs can be used to retrieve additional information at the subtest level.

Influence of Item, Person, and Test Characteristics

Levine & Drasgow (1982, 1983) investigated (1) the influence of using estimated item parameters instead of true parameters on the detection rate of l_0, and (2) the influence of the presence of misfitting item-score patterns on item parameter estimates and detection rates. Response vectors were simulated using the 3PLM and estimated item parameters from a previous calibration study of the Scholastic Aptitude Test (Verbal). Misfitting item-score vectors were simulated by randomly selecting 20% of the item scores (0s and 1s) from each vector and changing them with a probability of .20 (1s became 0s and 0s became 1s). They concluded that the detection rate of l_0 was not seriously affected by the estimated item parameters or the presence of misfitting item-score patterns. Kogut (1987), however, concluded from his simulation study that, as a result of the presence of deviant item-score patterns, the power of l_z and M was seriously reduced. Possible explanations for the different results from these studies are the different statistics used and the different numbers of simulated item-score patterns. In Levine and Drasgow, there were 6–7% misfitting response vectors, whereas there were 20% in the Kogut study. The higher percentage of misfitting item-score patterns might have reduced the power. Furthermore, the type of misfitting item-score vectors also might have been responsible for reduced power.

Reise & Due (1991) found that longer tests and a larger spread of item difficulties resulted in higher detection rates for l_z. They simulated item scores with lower Fisher information for estimating θ than was predicted by IRT model parameters. Different levels of the α parameter (which is related to item information; Hambleton & Swaminathan, 1985, p. 105) were used. Test length was 7, 21, 35, or 49 items, and the spread varied in the δs and γs. Reise and Due concluded that test length, δ spread, and γ each affected the detection rate of l_z. They found that, in general, longer tests, larger spread of δ, and lower γ values resulted in higher detection rates. Furthermore, l_z obtained its lowest detection rate for low θs.

Parsons (1983) investigated the effectiveness of a transformed version of l_0 for detecting simulated misfit on the Job Descriptive Index, which measures satisfaction with multiple facets of a job. Data were generated using the 2PLM and estimated item parameters from an empirical calibration sample. Twenty of the 60 items were selected, and scores were generated with a .30 probability of obtaining the correct response. Results indicated that higher detection rates were obtained at higher θs. The explanation was that, for these simulees, more item scores were changed. Furthermore, it was found that the variance of the NC score was lower for misfitting than for model-fitting patterns, because misfitting item scores were probably uncorrelated with each other, thus reducing the variance of the NC score compared to the variance of the NC score on a set of correlated items.

Smith (1985) compared robust estimators with person-fit statistics. Robust estimators correct for unexpected responses and weight them less to obtain a representative $\hat{\theta}$. Smith concluded that it is better to use person-fit analysis, because the robust estimators introduce a bias into θ estimation.

Reise (1995) investigated the detection rate of l_z as a function of true θ and several θ estimates: MLEs, expected a posteriori (EAP) estimates, and biweight (BIW) estimates. To estimate θ, datasets were simulated based on the estimated item parameters of four personality scales that fit the 2PLM. Reise found that using true θ consistently resulted in the highest detection rate for l_z. The detection rate differed among the three estimation methods, but the differences depended on the type of test,

θ level, and percentage of misfitting responses. Reise found higher detection rates for scales with a larger spread in item difficulties. BIW estimation typically resulted in a somewhat higher detection rate than EAP and MLE.

Meijer & Nering (1997) investigated the detection rate of l_z using MLE, EAP, and BIW, and also the bias in $\hat{\theta}$ as a function of different types of misfitting behavior. They found that the presence of misfitting item-score patterns influenced the bias in $\hat{\theta}$, and this depended heavily on the type of misfit and the θ level. The BIW scoring method reduced the bias in $\hat{\theta}$ and improved the detection rate relative to MLE and EAP for examinees located at both extremes of the θ continuum.

Application of Person-Fit Statistics to Empirical Data

Birenbaum (1985) compared the effectiveness of *ECI1*, *ECI2*, *ECI4*, their standardized versions, l_0, l_z, and U in distinguishing among empirical item-score patterns: item scores of an uncooperative group, item scores of a cooperative group, and randomly generated item scores. The first two groups were distinguished from each other on the basis of (1) motivation to take a test (rated by a test administrator) and (2) whether the student wrote his/her name on the test answer sheet. The test was administered only for research and development purposes. Except for U, Birenbaum found significant differences in the mean value of the statistics among the three groups. The correlations among the standardized indices were high (.90), but l_0 and U had a low correlation (.10). Most statistics had low correlations with NC scores (.13–.22). Curvilinearity between person-fit statistics and NC scores was not rejected for any of the unstandardized *ECI*s. Largest curvilinearity was detected for l_0, indicating that it yielded the most inflated values at both extremes of the θ scale.

In another study, Birenbaum (1986) investigated the relationships among $ECI1_z$, $ECI2_z$, $ECI4_z$, and l_z, the scores on an anxiety scale and a lie scale of the MMPI, and a general ability test. Birenbaum predicted that a sample of examinees with low anxiety scores and high lie scores would have less appropriate item-score patterns on the test than would low-anxiety examinees with low lie-scale scores. These predictions were based on the assumption that persons with high lie scores typically have the desire to deliberately impress the assessor by saying that they have low anxiety, but are unable to conceal the effect of their anxiety on the cognitive reasoning test scores. High correlations were found among the person-fit statistics (.97–.99). Low correlations were found between the scores on the lie scale and the fit statistics (mean = .10) and between the scores on the anxiety scale and the fit statistics (mean = .14). Scores on the ability test correlated .50 with the fit indices. There was a significant difference between the mean scores of the person-fit statistics for two groups: examinees with low anxiety and high lie scores were more misfitting than examinees with low anxiety and low lie scores.

Hoijtink (1987) investigated the effect of misfitting item-score patterns on item fit to the Rasch model. Item-score patterns were from two empirical datasets of a questionnaire measuring neurological and ophthalmic skills for general practitioners. Misfitting item-score patterns were removed from the dataset to determine whether this resulted in a better fit of misfitting *items* to the Rasch model. To minimize the danger of adapting the data to the model, item-score patterns were removed only under the condition that (1) they were classified as misfitting using both the original and improved item estimates, and (2) the fit of the dataset as a whole would be improved after removing misfitting examinees. Hoijtink showed that removing misfitting item-score patterns resulted in a better fit to the model for some items. However, it could not be explained why some examinees answered the questionnaires in a deviant way, as was done in the Birenbaum (1985) study.

Phillips (1986) also investigated the effect of deleting misfitting item-score patterns on the fit of the Rasch model, estimated item parameters, and equipercentile equating results. It was found

that substantially more items fit the Rasch model after deleting misfitting item-score patterns (an increase of approximately 50%). The effect of removing misfitting items scores on the estimated item difficulty parameters was small. Equating results were similar for datasets with and without misfitting item-score patterns.

Rudner, Bracey, & Skaggs (1996) investigated the use of W in the 1990 NAEP Trial State Assessment. They found almost no examinees with extreme item-score patterns. Eliminating examinees with the worst fit did not result in meaningful differences in the mean NAEP scale scores between trimmed and untrimmed data.

Reise & Waller (1993) explored the use of l_z in personality measurement by analyzing empirical data from the Multidimensional Personality Questionnaire (Tellegen, 1982). They noted that it is difficult to distinguish persons not fitting the particular trait from misfit due to measurement error or faulty responding. To reduce the chances of misfit due to measurement error or faulty responding, they used unidimensional subscales and information from detection scales that identify inconsistent answering behavior. l_z was able to identify persons not responding according to the 2PLM who were not identified by inconsistency scales. However, the accuracy of the classification could not be evaluated, because it was unknown which persons really behaved in an unusual manner.

Zickar & Drasgow (1996) analyzed a dataset from a personality test consisting of item scores from examinees instructed either to respond honestly or to fake the answers to convey a favorable impression. They found that optimal person-fit statistics classified a higher number of faking examinees than did a social desirability scale. The detection rates, however, were low (mostly between .10 and .30).

Molenaar & Hoijtink (1996) investigated the use of M (Equation 33) in the Rasch model for a test in which four- to seven-year-old children indicated which of three pictures presented was consistent with an item. Each picture had a number of balls and stars, colored white and black. They identified patterns with low probability of exceedance. For example, ordering the items from easy to difficult, an 11-item test had the response pattern (00000010011), which had a significance probability of .002. They concluded that this pattern was a candidate for closer inspection.

Schmitt, Chan, Sacco, McFarland, & Jennings (1999) used l_{zm} for investigating the relationship between test-taking motivation and conscientiousness in personality and cognitive tests. The relationship between gender and race (African-Americans and Whites) and l_{zm} also was investigated. Based on literature about the visual inspection of item-score patterns, males and African-Americans were expected to produce more irregular item scores than females and Whites. These differences might be explained by different test-taking motivation across race and gender subgroups. Schmitt et al. found that test-taking motivation correlated .26 with l_{zm} for personality tests and .12 for cognitive tests. Conscientiousness correlated .34 with l_{zm} for personality tests. Males had lower l_{zm} mean values than females (indicating misfit) on both cognitive and personality tests, but controlling for conscientiousness reduced or eliminated this association. This result was not observed for test-taking motivation. For cognitive tests, African-Americans obtained lower mean person-fit scores than Whites, but for personality data there was no difference. These scores were not related to conscientiousness or test-taking motivation. Schmitt et al. concluded that carelessness might explain the misfit of males, whereas a more general trait-like measure might influence reactions of African-Americans. Meijer & van Krimpen-Stoop (2001), using achievement test data, also found smaller mean values of l_z for men than for women and smaller l_z values for African-Americans than for Whites.

Validity and Misfitting Response Behavior

The relationship between deviant response behavior and decision making was studied by Drasgow & Guertler (1987). They suggested that overestimating $\hat{\theta}$ might result in selecting persons

unable to perform adequately on a job; similarly, underestimating $\hat{\theta}$ might be expensive for an organization due to extra selection efforts needed. They presented a utility theory approach to the use of person-fit statistics in practical settings. The approach requires the distribution of a statistic in samples with model-fitting and misfitting item-score patterns. Using the probabilities of score patterns under these distributions, the utility can be estimated and the critical value of a statistic can be determined. Drasgow and Guertler illustrated their approach using empirical data from the Armed Services Vocational Aptitude Battery. They concluded that combining utilities with classification on the basis of a person-fit statistic is complicated and involves many subjective judgments.

Schmitt, Cortina, & Whitney (1993) investigated whether misfitting item-score patterns distorted criterion-related validity estimates and estimates of the relationship between trait levels and performance constructs. Using l_z, the 3PLM, and four empirical datasets, they found little or no improvement of the correlation between a predictor and a criterion when misfitting item-score patterns were removed from the data. However, a hierarchical regression analysis in which the criterion scores were regressed onto (1) the predictor scores, (2) group membership based on l_z scores (model-fitting or misfitting), and (3) their cross-products, showed a significant interaction term for some datasets, implying that l_z scores might improve prediction.

Meijer (1997) used simulated data to investigate the relationship between misfit and test-score validity. He concluded that misfitting item-score patterns can influence the validity of a test if the type of misfit is severe, the correlation between the predictor and criterion scores is .3 or .4, and the percentage of misfitting item-score patterns is relatively high (at least .15 or higher). However, using l_z for removing misfitting item-score patterns from a predictor test had little impact on criterion-related validity. These results confirmed the results found by Schmitt et al. (1993) and can partly be explained by the less-than-perfect detection rate. In the most favorable case, approximately 40% of the misfitting item-score patterns remained in the sample.

Meijer (1998) used *ZU3* to identify persons with unexpected item scores on empirical selection data. In general, persons with inconsistent item scores were less predictable than persons with consistent item scores. Persons with both lower and higher criterion scores than predicted could be identified.

Discussion and Conclusions

Choosing a Statistic

For short or moderate test length (e.g., 10–60 items) and assuming a standard normal distribution of a person-fit statistic, the nominal and empirical Type I error rates are not in agreement for most statistics, because $\hat{\theta}$ is used instead of θ. Recently, Snijders (1998) proposed statistical theory for correcting the bias caused by using $\hat{\theta}$ rather than θ. Sound person-fit methods have been derived for the Rasch model, but because it is restrictive with respect to empirical data, the use of these statistics also is restricted.

Before using a particular person-fit statistic, a researcher should investigate possible threats to the fit of individual item-score patterns. If violations of local independence are expected, a method proposed by Klauer (1991) might be used instead of, for example, Molenaar & Hoijtink's (1990) M. Tests against a specific alternative generally are more powerful than general statistics, and the type of deviance is easier to interpret. Statistics like M are helpful when the threats are unknown. UB (Equation 24) and UW (Equation 25) or the PRF can be used as diagnostic tools to test whether item-score patterns on a priori specified subtests fit an IRT model.

For some person-fit statistics (e.g., l_z), only deviations against the model are tested, resulting in interpretation problems. For example, item-score patterns not fitting the Rasch model might be

better described by the 3PLM. If a model does not fit the data, other explanations are possible. It is often difficult to substantively distinguish different types of item-score patterns and/or to obtain additional information using background variables. Testing against specific alternatives might be a better strategy.

Almost all statistics are of the form given in Equation 17, but with different weights. The use of a statistic depends on which model is used. Using the Rasch model, Molenaar & Hoijtink's (1990) M is a good choice. M should be preferred over l_z or ZW, because the critical values for M are more accurate. M is available in the computer program RSP (Glas & Ellis, 1993); practitioners can easily add person-fit values to their datasets.

Residual-based statistics (e.g., Equations 24 and 25) do not reflect the probability of ordering of the score patterns, because $1/\{P_g(\theta)[1 - P_g(\theta)]\}$ is not an increasing function in $P_g(\theta)$.

In a nonparametric context, *ZU3* might be preferred over the other fit statistics (e.g., C^*), because it is also an increasing function of the probability of the item-score pattern. Moreover, the distribution of *ZU3* is known to be standard normal conditional on the NC score. However, Emons, Meijer, & Sijtsma (in press) showed that the empirical distribution is not in agreement with the theoretical distribution when nonparametric IRT models are used.

Improving Measurement Practice

The objective of person-fit measurement is to detect item-score patterns that are improbable given an IRT model or given the other patterns in a sample. Thus, person-fit statistics must be sensitive to misfitting item-score patterns. Research has shown that detection rates are highly dependent on the (1) type of misfitting response behavior, (2) θ value, and (3) test length. When item-score patterns do not fit an IRT model, high detection rates can be obtained for extreme θs, even when Type I errors are low (e.g., .001). This is because deviations from model predictions tend to be larger for extreme θs than for moderate θs. The bias in $\hat{\theta}$ tends to be larger for extreme θs than for moderate θs (Meijer & Nering, 1997).

Relatively few studies have investigated the usefulness of person-fit statistics for analyzing empirical data. The few studies that exist have found some evidence that groups of persons with a priori known characteristics, such as low test-taking motivation, produced deviant item-score patterns that are unlikely given an IRT model. However, the usefulness of person-fit data depends on the degree of misfitting response behavior. Although additional empirical research is needed (Reise & Flannery, 1996; Rudner et al., 1996), more empirical studies will not necessarily demonstrate whether person-fit statistics can be helpful in improving measurement practice. Empirical studies can illustrate their use; however, whether person-fit statistics can help a researcher in practice depends on the context in which research takes place.

Smith (1985) mentioned four actions that could be taken when an item-score pattern is classified as misfitting: (1) report several θ estimates (rather than just one) for an examinee based on subtests that are in agreement with the model, (2) modify the item-score pattern (e.g., eliminate the unreached items at the end) and re-estimate θ, (3) do not report the θ estimate and retest the examinee, or (4) decide that the error is small enough for the impact on $\hat{\theta}$ to be marginal. The latter action can be based on comparing the error introduced by misfitting item-score patterns and the standard error associated with each θ estimate. Which of these actions is taken depends on the context in which testing takes place. The usefulness of person-fit statistics thus also depends heavily on the application for which it is intended.

Suggestions for Future Research

1. Research is needed for methods that compensate for the use of $\hat{\theta}$. Snijders (1998) proposed a correction for the standard error of a group of person-fit statistics. However, the skewness

and kurtosis of the sampling distributions also should be taken into account, especially when the nominal Type I levels are small (i.e., when nominal and empirical Type I error levels are not in agreement; van Krimpen-Stoop & Meijer, 1999).

2. The use of new statistics (e.g., Bradlow et al., 1998; McLeod & Lewis, 1999; van Krimpen-Stoop & Meijer, 2000) within the context of CAT should be more thoroughly investigated.
3. Studies are needed analyzing empirical data together with background variables to obtain extra information about the type of misfit (Meijer & de Leeuw, 1993). Research is also needed to distinguish between examinees with item-score patterns for whom an inappropriate IRT model is used and those whose item-score patterns can be explained using additional information. Reise & Flannery (1996) mentioned the application of person-fit research in cross-cultural studies to investigate the scalability of examinees with different ethnic backgrounds.
4. Few person-fit statistics with known statistical properties exist within nonparametric IRT modeling that test an item-score pattern against model assumptions. More research is needed here to obtain sound statistical methods.
5. Testing against a specified alternative might be a solution to the relatively low power of person-fit statistics in detecting misfit. More information is needed about the influence of misfitting response behaviors on test scores.
6. Misfitting item-score patterns might be more easily interpreted using external frames of reference (e.g., evaluating the fit of an item-score pattern using the item difficulties determined in a well-defined group of examinees or on the basis of a cognitive theory; Embretson, 1993). Latent class analysis might be useful, if classes with a specific type of misfitting response behavior can be incorporated into the model (see van den Wittenboer, Hox, & de Leeuw, 1997, 2000, for an example).
7. The PRF also might be used to enhance the interpretation of misfitting item-score patterns (Sijtsma & Meijer, in press). Because a plot of the observed and expected response functions immediately clarifies which groups of observed responses disagree with the expected responses, researchers might more easily hypothesize the explanation of the misfitting item-score patterns. Reise (2000) described the application of multilevel analysis to the analysis of person fit using the PRF.
8. Although most studies using person-fit statistics are based on IRT models, recently Reise & Widaman (1999) explored the use of person-fit statistics within covariance structure models. Reise and Widaman proposed an index to assess the contribution of an examinee's fit to a covariance structure model. A first comparison of the l_z statistic and this index using empirical data resulted, however, in a correlation of almost 0—indicating that both indices classified different examinees as misfitting. More research with this approach is indicated.

References

Aguinis, H., & Stone-Romero, E. F. (1997). Methodological artifacts in moderated multiple regression and their effects on statistical power. *Journal of Applied Psychology, 82,* 192–206.

Bedrick, E. J. (1997). Approximating the conditional distribution of person-fit indexes for checking the Rasch model. *Psychometrika, 62,* 191–199.

Birenbaum, M. (1985). Comparing the effectiveness of several IRT based appropriateness measures in detecting unusual response patterns. *Educational and Psychological Measurement, 45,* 523–534.

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

Bradlow, E. T., Weiss, R. E., & Cho, M. (1998). Bayesian identification of outliers in computerized adaptive testing. *Journal of the American Statistical Association, 93,* 910–919.

Cliff, N. (1983). Evaluating Guttman scales: Some old and new thoughts. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement: A festschrift for Frederic M. Lord* (pp. 283–300). Hillsdale NJ: Erlbaum.

Cliff, N. (1996). *Ordinal methods for behavioral data analysis.* Mahwah NJ: Erlbaum.

Donlon, T. F., & Fischer, F. E. (1968). An index of an individual's agreement with group determined item difficulties. *Educational and Psychological Measurement, 28,* 105–113.

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

Drasgow, F., & Guertler, E. (1987). A decision-theoretic approach to the use of appropriateness measurement for detecting invalid test and scale scores. *Journal of Applied Psychology, 72,* 10–18.

Drasgow, F., & Levine, M. V. (1986). Optimal detection of certain forms of inappropriate test scores. *Applied Psychological Measurement, 10,* 59–67.

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.

Drasgow, F., Levine, M. V., & Zickar, M. J. (1996). Optimal identification of mismeasured individuals. *Applied Measurement in Education, 9,* 47–64.

Ellis, J. L., & van den Wollenberg, A. L. (1993). Local homogeneity in latent trait models. A characterization of the homogeneous monotone IRT model. *Psychometrika, 58,* 417–429.

Embretson, S. E. (1993). Psychometric models for learning and cognitive processes. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 125–150). Hillsdale NJ: Erlbaum.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah NJ: Erlbaum.

Emons, W. H. M., Meijer, R. R., & Sijtsma, K. (in press). Comparing simulated and theoretical sampling distributions of the *U3* person-fit statistic. *Applied Psychological Measurement.*

Glas, C. A. W., & Ellis, J. L. (1993). *User's manual: RSP.* Groningen, The Netherlands: ProGAMMA.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9,* 139–150.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A Claussen (Eds.), *Measurement and prediction* (pp. 60-90). Princeton NJ: Princeton University Press.

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

Harnisch, D. L. (1983). Item response patterns: Applications for educational practice. *Journal of Educational Measurement, 20,* 191–205.

Harnisch, D. L., & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement, 18,* 133–146.

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

Hoijtink, H. (1987). Rasch schaal constructie met behulp van een passingsindex voor personen [Rasch scale construction using a person-fit index]. *Kwantitatieve Methoden, 25,* 101–110.

Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika, 55,* 577–601.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory.* Homewood IL: Dow Jones-Irwin.

Jaeger, R. M. (1988). Use and effect of caution indices in detecting aberrant patterns of standard-setting judgments. *Applied Measurement in Education, 1,* 17–31.

Kane, M. T., & Brennan, R. L. (1980). Agreement coefficients as indices of dependability for domain-referenced tests. *Applied Psychological Measurement, 4,* 105–126.

Kendall, M. G. (1970). *Rank correlation methods* (4th ed.). London: Griffin.

Klauer, K. C. (1991). An exact and optimal standardized person test for assessing consistency with the Rasch model. *Psychometrika, 56,* 535–547.

Klauer, K. C. (1995). The assessment of person fit. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models. Foundations, recent developments, and applications* (pp. 97–110). New York: Springer-Verlag.

Klauer, K. C., & Rettig, K. (1990). An approximately standardized person test for assessing consistency with a latent trait model. *British Journal of Mathematical and Statistical Psychology, 43,* 193–206.

Kogut, J. (1986). *Review of IRT-based indices for detecting and diagnosing aberrant response patterns* (Research Report No. 86-4). Enschede, The Netherlands: University of Twente.

Kogut, J. (1987). *Detecting aberrant response patterns in the Rasch model* (Research Report No. 87-3). Enschede, The Netherlands: University of Twente.

Kogut, J. (1988). *Asymptotic distribution of a person-fit statistic* (Research Report No. 88-13). Enschede, The Netherlands: University of Twente.

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

Levine, M. V., & Drasgow, F. (1983). Appropriateness measurement: Validating studies and variable ability models. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 109–131). New York: Academic Press.

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

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

Li, M. F., & Olejnik, S. (1997). The power of Rasch person-fit statistics in detecting unusual response patterns. *Applied Psychological Measurement, 21,* 215–231.

Lindgren, B. W. (1993). *Statistical theory.* London: Chapman and Hall.

Liou, M. (1993). Exact person tests for assessing model-data fit in the Rasch model. *Applied Psychological Measurement, 17,* 187–195.

Liou, M., & Chang, C. H. (1992). Constructing the exact significance level for a person fit statistic. *Psychometrika, 57,* 169–181.

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

Lumsden, J. (1977). Person reliability. *Applied Psychological Measurement, 1,* 477–482.

Lumsden, J. (1978). Tests are perfectly reliable. *British Journal of Mathematical and Statistical Psychology, 31,* 19–26.

McLeod, L. D., & Lewis, C. (1998, April). *A Bayesian approach to detection of item preknowledge in a CAT.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego CA.

McLeod, L. D., & Lewis, C. (1999). Detecting item memorization in the CAT environment. *Applied Psychological Measurement, 23,* 147–160.

Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement, 18,* 311–314.

Meijer, R. R. (1996). The influence of the presence of deviant item score patterns on the power of a person-fit statistic. *Applied Psychological Measurement, 20,* 141–154.

Meijer, R. R. (1997). Person fit and criterion-related validity: An extension of the Schmitt, Cortina, and Whitney study. *Applied Psychological Measurement, 21,* 99–113.

Meijer, R. R. (1998). Consistency of test behaviour and individual difference in precision of prediction. *Journal of Occupational and Organizational Psychology, 71,* 147–160.

Meijer, R. R., & de Leeuw, E. D. (1993). Person fit in survey research: The detection of respondents with unexpected response patterns (pp. 235–245). In J. H. Oud & R. A. W. van Blokland-Vogelesang (Eds.), *Advances in longitudinal and multivariate analysis in the behavioral sciences.* Nijmegen, The Netherlands: ITS.

Meijer, R. R., Molenaar, I. W., & Sijtsma, K. (1994). Influence of test and person characteristics on nonparametric appropriateness measurement. *Applied Psychological Measurement, 18,* 111–120.

Meijer, R. R., & Nering, M. L. (1997). Trait level estimation for nonfitting response vectors. *Applied Psychological Measurement, 21,* 321–336.

Meijer, R. R., & Nering, M. L. (1999). Computerized adaptive testing: Overview and introduction. *Applied Psychological Measurement, 23,* 187–194.

Meijer, R. R., & Sijtsma, K. (1995). Detection of aberrant item score patterns: A review and new developments. *Applied Measurement in Education, 8,* 261–272.

Meijer, R. R., & van Krimpen-Stoop, E. M. L. A. (2001). Person fit across subgroups: An achievement testing example. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 377–390). New York: Springer-Verlag.

Miller, M. D. (1986). Time allocation and patterns of item response. *Journal of Educational Measurement, 23,* 147–156.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17,* 297–334.

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

Molenaar, I. W., & Hoijtink, H. (1996). Person-fit and the Rasch model, with an application to knowledge of logical quantors. *Applied Measurement in Education, 9,* 27–45.

Nering, M. L. (1995). The distribution of person fit using true and estimated person parameters. *Applied Psychological Measurement, 19,* 121–129.

Nering, M. L. (1997). The distribution of indexes of person-fit within the computerized adaptive testing environment. *Applied Psychological Measurement, 21,* 115–127.

Nering, M. L., & Meijer, R. R. (1998). A comparison of the person response function and the l_z person-

fit statistic. *Applied Psychological Measurement, 22,* 53–69.

Noonan, B. W., Boss, M. W., & Gessaroli, M. E. (1992). The effect of test length and IRT model on the distribution and stability of three appropriateness indexes. *Applied Psychological Measurement, 16,* 345–352.

Nunnally, J. C. (1978). *Psychometric theory.* New York: McGraw-Hill.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika, 41,* 100–115.

Parsons, C. K. (1983). The identification of people for whom Job Descriptive Index scores are inappropriate. *Organizational Behavior and Human Performance, 33,* 365–393.

Phillips, S. E. (1986). The effects of deletion of misfitting persons on vertical equating via the Rasch model. *Journal of Educational Measurement, 23,* 107–118.

Rasch, G. (1960/80). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Danish Institute for Educational Research. Expanded edition (1980), with foreword and afterword by B. D. Wright. Chicago: University of Chicago Press.

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. (1995). Scoring method and the detection of person misfit in a personality assessment context. *Applied Psychological Measurement, 19,* 213–229.

Reise, S. P. (2000). Using multilevel logistic regression to evaluate person-fit in IRT models. *Multivariate Behavioral Research, 35,* 543–570.

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., & Flannery, W. P. (1996). Assessing person-fit measurement of typical performance applications. *Applied Measurement in Education, 9,* 9–26.

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

Reise, S. P., & Widaman, K. F. (1999). Assessing the fit of measurement models at the individual level: A comparison of item response theory and covariance structure models. *Psychological Methods, 4,* 3–21.

Rogers, H. J., & Hattie, J. A. (1987). A monte carlo investigation of several person and item fit statistics for item response models. *Applied Psychological Measurement, 11,* 47–57.

Rudner, L. M. (1983). Individual assessment accuracy. *Journal of Educational Measurement, 20,* 207–219.

Rudner, L. M., Bracey, G., & Skaggs, G. (1996). The use of a person-fit statistic with one high quality achievement test. *Applied Measurement in Education, 9,* 91–109.

Sato, T. (1975). *The construction and interpretation of S-P tables.* Tokyo: Meiji Tosho.

Schmitt, N., Chan, D., Sacco, J. M., McFarland, L. A., & Jennings, D. (1999). Correlates of person fit and effect of person fit on test validity. *Applied Psychological Measurement, 23,* 41–53.

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

Sijtsma, K. (1986). A coefficient of deviance of response patterns. *Kwantitatieve Methoden, 7,* 131–145.

Sijtsma, K., & Meijer, R. R. (1992). A method for investigating the intersection of item response functions in Mokken's nonparametric IRT model. *Applied Psychological Measurement, 16,* 149–157.

Sijtsma, K., & Meijer, R. R. (in press). The person response function as a tool in person-fit research. *Psychometrika.*

Smith, R. M. (1985). A comparison of Rasch person analysis and robust estimators. *Educational and Psychological Measurement, 45,* 433–444.

Smith, R. M. (1986). Person fit in the Rasch model. *Educational and Psychological Measurement, 46,* 359–372.

Snijders, T. (in press). Asymptotic distribution of person-fit statistics with estimated person parameter. *Psychometrika.*

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3,* 271–295.

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.

Tatsuoka, K. K., & Tatsuoka, M. M. (1982). Detection of aberrant response patterns and their effect on dimensionality. *Journal of Educational Statistics, 7,* 215–231.

Tatsuoka, K. K., & Tatsuoka, M. M. (1983). Spotting erroneous rules of operation by the individual consistency index. *Journal of Educational Measurement, 20,* 221–230.

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

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

van den Wittenboer, G., Hox, J. J., & de Leeuw, E. D. (1997). Aberrant response patterns in elderly respondents: Latent class analysis of respondent scalability (pp. 155–162). In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences.* Munster: Waxman Verlag.

van den Wittenboer, G., Hox, J. J., & de Leeuw, E. D. (2000). Latent class analysis of respondent scalability. *Quality and Quantity, 34,* 177–191.

van der Flier, H. (1977). Environmental factors and deviant response patterns (pp. 30–35). In Y. P. Poortinga (Ed.), *Basic problems in cross-cultural psychology.* Amsterdam: Swets and Zeitlinger.

van der Flier, H. (1980). *Vergelijkbaarheid van individuele testprestaties* [Comparability of individual test performance]. Lisse: Swets and Zeitlinger.

van der Flier, H. (1982). Deviant response patterns and comparability of test scores. *Journal of Cross-Cultural Psychology, 13,* 267–298.

van der Linden, W. J., & Hambleton, R. K. (Eds.) (1997). *Handbook of modern item response theory.* New York: Springer-Verlag.

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (1999). Simulating the null distribution of person-fit statistics for conventional and adaptive tests. *Applied Psychological Measurement, 23,* 327–345.

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2000). Detecting person-misfit in adaptive testing using statistical process control techniques. In W. J. van der Linden and C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 201–219). Boston: Kluwer.

Weiss, D. J. (1973). *The stratified adaptive computerized ability test* (Research Report No. 73-3). Minneapolis MN: University of Minnesota, Department of Psychology.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 4,* 273–285.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* Chicago: MESA Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design. Rasch measurement.* Chicago: Mesa Press.

Zickar, M. J., & Drasgow, F. (1996). Detecting faking on a personality instrument using appropriateness measurement. *Applied Psychological Measurement, 20,* 71–87.

Acknowledgments

The authors thank Edith M. L. A. van Krimpen-Stoop for her comments on an earlier draft of this paper, as well as two anonymous reviewers and the editor for their help with improving this paper.

Author's Address

Send requests for reprints or further information to Rob R. Meijer, University of Twente, TO/OMD, P. O. Box 217, 7500 AE Enschede, The Netherlands. Email: meijer@edte.utwente.nl.