DOCUMENT RESUME

ED 434 148 TM 030 112

AUTHOR Meijer, Rob R.; Sijtsma, Klaas
TITLE A Review of Methods for Evaluating the Fit of Item Score Patterns on a Test. Research Report 99-01.
INSTITUTION Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.
PUB DATE 1999-00-00
NOTE 55p.
AVAILABLE FROM Faculty of Educational Science and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
PUB TYPE Reports - Descriptive (141)
EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Evaluation Methods; *Goodness of Fit; *Item Response Theory; Personality Measures; *Scores; Test Construction; *Test Items
IDENTIFIERS *Person Fit Measures

ABSTRACT

Methods are discussed that can be used to investigate the fit of an item score pattern to a test model. Model-based tests and personality inventories are administered to more than 100 million people a year and, as a result, individual fit is of great concern. Item Response Theory (IRT) modeling and person-fit statistics that are formulated in the context of IRT take a prominent place in the literature. Person-fit statistics are extensively discussed in this paper. Also, methods formulated outside the IRT context and methods to investigate particular types of response behavior are discussed. The aim of this paper is to give the researcher an idea of the possibilities in this research area by emphasizing the similarities of most person-fit methods and by discussing the pros and cons of the methods. (Contains 98 references and a list of University of Twente research reports.) (Author/SLD)

A Review of Methods for Evaluating the Fit of Item Score Patterns on a Test

Research Report 99-01

Rob R. Meijer, University of Twente
Klaas Sijtsma, Tilburg University

faculty of
EDUCATIONAL SCIENCE AND TECHNOLOGY

University of Twente

Department of
Educational Measurement and Data Analysis

A Review of Methods for Evaluating the Fit of Item Score Patterns on a Test

Rob R. Meijer
University of Twente

Klaas Sijtsma
Tilburg University

Methods are discussed that can be used to investigate the fit of an item score pattern to a test model. Model-based tests and personality inventories are administered to more than 100 million people a year and, as a result, individual fit is of great concern. Item response theory (IRT) modeling and person-fit statistics that are formulated in the context of IRT take a prominent place in the literature. Person-fit statistics are extensively discussed in this paper. Also, methods formulated outside the IRT context and methods to investigate particular types of response behavior are discussed. The aim of this paper is to give the researcher an idea of the possibilities in this research area by emphasizing the similarities of most person-fit methods and by discussing the pro's and con's of the methods.

Index terms: answer copying, appropriateness measurement, item response theory, person-fit measurement, test theory.

Inaccuracy of measurement is central to measurement theorists, measurement practitioners and educational policy makers. Since the beginning of standardized testing, inaccuracy of measurement has received widespread attention. Examples are reliability theory and methods for estimating reliability (Gulliksen, 1950; Lord & Novick, 1968; Spearman, 1910), statistics for comparing groups with respect to the probability of correctly answering an item (differential item functioning; e.g., Millsap & Everson, 1993), and differential prediction of subgroups of persons using moderated multiple regression analysis (e.g., Aguinis & Stone-Romero, 1997). In this review, we will give attention to research methods for determining the fit of individual item score patterns to a test model.

Researchers always have shown interest in obtaining additional information to the total score by studying patterns of individual item scores (e.g., Nunnally, 1978). Discriminant analysis and cluster analysis have been used to cluster similar types of score patterns for testing the hypothesis that a priori hypothesized groups can be distinguished (discriminant analysis) or to discover in an exploratory sense groups that have similar response patterns (cluster analysis). Both methods concentrate on groups, not on individual persons. In this review, methods are discussed that provide information at the individual level.

In the past two decades, important contributions to assessing individual test performance arose from item response theory (IRT); these contributions are summarized as person-fit research and will be discussed extensively in this review. In most person-fit research the fit of a score pattern to an IRT model is investigated. However, because person-fit research (although not under that name) has also been conducted without IRT modeling, approaches outside the IRT framework will also be discussed. Furthermore, related research with respect to answer copying statistics will be discussed because there are interesting relations between these statistics and person-fit measures.

The aim of this review is to present and discuss methods that can be used to detect nonfitting item score patterns. As such, this review can be considered an extension of Chapter 4 of the book of Hulin, Drasgow, and Parsons (1983) and Kogut (1986), in which overviews were given of such methods. Person-fit research has shown to be attractive for many researchers, which is corroborated by a proliferation of research articles in this area. A review of the present state of affairs thus seems to be justified. Moreover, this review is much more comprehensive than a recent review by Meijer and Sijtsma (1995)

which provided a general discussion of person-fit research.

Person Fit or Appropriateness Measurement ?

Methods that evaluate the fit of the individual test performance to an IRT model are usually referred to as appropriateness measurement methods or person-fit methods. Levine and Drasgow (1983, p. 110) seem to prefer the term appropriateness measurement for methods that "recognize inappropriate test scores". They furthermore state (Levine & Drasgow, 1983, p. 110) that "appropriateness measurement is only incidentally concerned with questions of person fit, the goodness of fit of a person's data to a test model (...)" and also " because all tractable models are inaccurate (...) a model and a measure of model fit cannot be considered useful for appropriateness measurement until they have been shown to effectively classify appropriate and inappropriate test scores" . Although we fully agree that all models are inaccurate for describing individual response behavior, we think that in practice appropriateness measurement and person-fit measurement are one and the same because most methods (and especially the methods that are used by Drasgow, Levine and colleagues) describe response behavior based on some type of test model. This implies that the appropriateness of a test score is defined on the basis of the (non)fitting of an item score pattern to a test model. Person fit is a more general terminology than appropriateness measurement and we will use "person fit" to indicate statistical methods for evaluating the fit of individual test performance to an IRT model or to other item score patterns in a sample; the term appropriateness measurement is rather vague in this respect.

Rationale for Person-Fit Research

As a measure of a person's ability level, the total score (or the trait level estimate) may be inadequate. For example, a person may guess some of the correct answers to multiple-choice items, thus raising his/her total score on the test by luck but not by ability, or an examinee not familiar with the test format may due to this unfamiliarity obtain a lower score than expected on the basis of his/her ability level (e.g., Wright & Stone, 1979, pp. 165-190). Inaccurate measurement of the trait level may also be caused by sleeping behavior (inaccurately answering the first questions in a test as a result of, for example, problems of getting started), cheating behavior (copying the correct answers of another

examinee), and plodding behavior (working very slowly and methodically and, as a result, generating item score patterns which are too good to be true given the stochastic nature of a person's response behavior as assumed by most IRT models; see e.g., Ellis & van den Wollenberg, 1993; Holland, 1990).

It is important to realize that not all types of aberrant behavior will affect individual test scores. For example, a person may guess the correct answers to some of the items but also guess wrong on some of the other items, and as the result of the stochastic nature this guessing process may not result in substantially different test scores under most IRT models to be discussed below. Whether aberrant behavior will lead to nonfitting item score patterns depends on numerous factors such as the type and the amount of aberrant behavior.

Furthermore, it should be noted that although all methods discussed in this paper can be used to detect nonfitting item score patterns, several of these methods do not allow the mechanism that created the deviant item score patterns to be recovered. Other methods explicitly test against specific violations of a test model assumption or against particular types of deviant item score patterns. These methods may therefore facilitate the interpretation of nonfitting item score patterns.

Person-fit Methods Based on Group Characteristics

Statistics

Most person-fit statistics compare an individual's observed and expected item scores across the items from a test. The expected item scores are determined on the basis of an IRT model or on the basis of the observed item means in the sample. In this section, we deal with group-based statistics. In the next section we discuss IRT-based person-fit statistics.

To demonstrate the similarity between several statistics, a general formula is used in which a particular choice of the weights (w) defines a particular person-fit statistic. Let n persons take a test consisting of k items and let π_g denote the proportion-correct score on item g that can be estimated from the sample by $\hat{\pi}_g = n_g/n$, where n_g is the number of 1 scores in the sample. Furthermore, let the items be ordered and numbered according to decreasing proportion-correct score (increasing item difficulty): $\pi_1 > \pi_2 > \ldots > \pi_k$,

and let the realization of a dichotomous (0,1) item score be denoted by $X_g = x_g$ (g=1, ..., k). Examinees are indexed i, with $i = 1, ..., n$. The number-correct score $X = r$ is the unweighted sum of item scores, $\sum_{g=1}^{k} X_g = r$. The general formula for group-based statistics is given by

$$G_i \equiv \frac{\sum_{g=1}^{r} w_g - \sum_{g=1}^{k} X_g w_g}{\sum_{g=1}^{r} w_g - \sum_{g=k-r+1}^{k} w_g} \quad (1)$$

To enhance interpretation of G_i, often person-fit statistics are normed against the range of possible values of G_i given the definition of w_i.

Person-fit statistics that are based on group characteristics compare an individual's item score pattern with the item score patterns of the other persons in the sample. Most person-fit statistics are a count of certain score patterns for item pairs and compare this count with the expectation under the deterministic Guttman (1944, 1950) model. Let θ be the latent trait known from IRT, and let δ be the location parameter which is measured on the scale θ. $P_g(\theta)$ is the conditional probability of giving a correct answer to item g. The Guttman model is defined by

$$\theta < \delta_g \Leftrightarrow P_g(\theta) = 0;$$

and

$$\theta \geq \delta_g \Leftrightarrow P_g(\theta) = 1.$$

The Guttman model thus 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 $g < h$. Such item score combinations (0,1) are called "errors" or "inversions". Item score patterns (1,0) are permitted, and are known as "Guttman patterns" or "conformal" patterns. A coefficient that has received some attention is the modified caution index (C_i^*) proposed by Harnisch and Linn (1981). C_i^* is a slight adaptation of the caution index (C_i; Sato, 1975). C_i^* can be obtained from Equation (1) by choosing $w_g = \pi_g$. C_i also is obtained by choosing $w_g = \pi_g$, and then multiplying $\sum_{g=k-r+1}^{k} w_g$ by r and the other terms by k. Both statistics weigh the item scores with the proportion-correct score

in the sample normed against the Guttman model. For example, it can easily be checked that C_i^* = 0 if an examinee with total score $X = r$ answers the r easiest items correctly and the $k - r$ most difficult items incorrectly; this means that an examinee's item score is in agreement with the Guttman (1950) model. Also note that C_i^*= 1 if the item score pattern equals the reversed Guttman pattern, thus indicating maximum aberrance. The lower bound of C_i also equals 0 when an item score pattern is in agreement with the Guttman model. However, C_i does not have a fixed upper bound and thus the values of C_i are more difficult to interpret than those of C_i^*.

Coefficients similar to C_i^* have been discussed by Donlon & Fischer (1968), van der Flier (1977), and Tatsuoka and Tatsuoka (1982, 1983). Donlon & Fischer (1968) proposed to use the personal point-biserial correlation (r_{pbis}) as a person-fit statistic; r_{pbis} is simply the correlation across all items between an examinee's binary item scores and the vector containing the sample frequencies of the item scores. Furthermore, they proposed the personal biserial correlation (r_{bis}), which is the personal point-biserial correlation under the assumption of a continuous normally distributed variable underlying each of the item responses. Van der Flier (1977) defined $U1$ as the number of Guttman errors normed against the maximum number of Guttman errors given $X = r$; this maximum equals $r(k - r)$.

Tatsuoka and Tatsuoka (1983) discussed the norm conformity index,

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

The numerator contains the number of Guttman conformal (1,0) pairs of item scores multiplied by 2. In the case of a reversed Guttman item score vector, the number of conformal (1,0) pairs equals 0 and, consequently, NCI_i = 1. In the case of a Guttman item score vector, the number of (1,0) pairs of item scores is $r(k - r)$ and, consequently, NCI_i = -1. Note that NCI_i is perfectly related to $U1$: $NCI_i = 1 - 2U1$. Tatsuoka and Tatsuoka (1983) also discussed the Individual Consistency Index (ICI_i), which is equivalent to NCI_i, but is determined for subgroups of items that require the same cognitive solution strategy. Thus, 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.

Kane and Brennan (1980) mentioned the agreement index, the disagreement index, and the dependability index that can be used as group-based person-fit statistics. The agreement index is defined as

$$A_i = \sum_{g=1}^{k} X_g \pi_g. \tag{3}$$

Let $A_i(\max)$ be the maximum value of A_i given the total score r. This maximum value is obtained if, given r, the item score pattern is a Guttman pattern; that is

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

The disagreement index is defined as

$$D_i = A_i(\max) - A_i, \tag{4}$$

and the dependability index is defined as

$$E_i = \frac{A_i}{A_i(\max)}. \tag{5}$$

Note that D_i equals the numerator of C_i^* (see Equation 1, taking $w_g = \pi_g$).

Sijtsma (1986; see also Sijtsma & Meijer, 1992) proposed a person-fit statistic denoted H_i^T. For a fixed set of k items, let β_i denote the expected proportion of items to which person i gives the correct response across locally independent repeated measurements. Let β_{ij} denote the expected proportion of items to which both persons i and j respond correctly. Then $\sigma_{ij} = \beta_{ij} - \beta_i \beta_j$ is the covariance between the scores of persons i and j. Now label examinees so that $i < j$ implies $\beta_i \leqslant \beta_j$; the maximum covariance between two examinees is then obtained when $\beta_{ij} = \beta_i$ and therefore $\sigma_{ij}^{\max} = \beta_i(1 - \beta_j)$. For one examinee in relation to $n - 1$ examinees,

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

The maximum value of H_i^T equals 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) = 0; and $H_i^T < 0$ if this average covariance is negative. Note that 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 .

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 (1) by choosing

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

To correct for dependence on the total score, $U3$ was standardized given $X = r$, and this standardized statistic is given by

$$ZU3 = \frac{U3 - E(U3)}{[Var(U3)]^{1/2}}. \tag{6}$$

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. To obtain $ZU3$, $E(U3)$ and $Var(U3)$ are needed. Note that for given $X = r$ all terms in Equation (1) are constant, except for $\sum_{g=1}^{k} X_g w_g$. Van der Flier (1982) derived expressions for $E(\sum_{g=1}^{k} X_g w_g)$ and $Var(\sum_{g=1}^{k} X_g w_g)$.

Research Using Group-Based Statistics

Several studies have used simulated data and empirical data for investigating the usefulness of group-based statistics to detect aberrant item score patterns.

Harnisch and Linn (1981) used empirical data from a reading test and from a math test to obtain the correlation between several statistics and also their correlation with the total score. The statistics were C_i, C_i^*, r_{pbis}, r_{bis}, A_i, D_i, E_i, and NCI_i . Harnisch and Linn (1981) found that for both tests, the correlations between almost all statistics were between .65 and .90, except for A_i which correlated approximately .40 with each of the other statistics. Most statistics correlated approximately .5 with the total score on both tests, except for C_i^* which correlated .20 (lowest) with the total score, and A_i which correlated .99 with the total score. Furthermore, Harnisch and Linn (1981) compared the average C_i^* scores across students for groups of students from different schools. They found sig-

nificant between-school differences that were attributed to instructional and curriculum differences.

Rudner (1983) used simulated data to compare r_{pbis}, r_{bis}, NCI_i, and C_i with several IRT-based person-fit statistics (U, W, and l, to be discussed below). High correlations ranging from .61 to .99 were found between the four group-based statistics. Two cases were distinguished in order to investigate the effectiveness of the statistics to detect aberrant item score patterns. In one case, for a minority of examinees several correct responses were randomly selected and then changed into incorrect responses, thus producing spuriously low number-correct scores. In the second case, some incorrect responses were changed into correct responses producing high number-correct scores. To identify whether the person-fit statistics could identify the altered item score patterns, Rudner (1983) checked whether the spuriously high or low scores were correctly classified as aberrant by the statistics. In general, the conclusion was that the effectiveness of detecting aberrant item score patterns increased with the number of altered items. For example, with an 11% change of incorrect responses into correct responses NCI_i produced a detection rate of .10, and with a 33% change NCI_i produced a detection rate of .20. Another conclusion was that for tests consisting of 45 items r_{bis} performed better than NCI_i and C_i, but for longer tests (80 items) the IRT-based statistic U_i performed best.

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. It was found that differences in time spend on a particular subject matter for which the test was intended resulted in different types of item score patterns and that in classes having a high C_i other topics were emphasized than in classes having a low C_i. Miller (1986) used the within-class standard deviation to interpret the mean C_i.

Tatsuoka and Tatsuoka (1983) used NCI_i to detect deviant item score patterns in an arithmetic test. They compared two groups of examinees. One group consisted of students who were far off the mastery level and who made many different kinds of errors. The other group consisted of students who were close to the mastery level and only made sophisticated errors. Because of these differences the item difficulties were different for the two groups. It was found that examinees who made only sophisticated errors and who were included in the group far away from the mastery stage, were classified as aberrants, but when these same examinees were included in the group which was close to mastery

these examinees were classified as normal. This empirical example illustrated that NCI_i obtains a relatively high value (indicating aberrance) if an examinee's item scores deviate from the item scores of a majority of examinees in the group. In the same study, ICI_i was used to identify examinees with inconsistent item score patterns on items that require similar cognitive skills.

Jaeger (1988) used C_i^* to identify judges whose patterns of item judgment were aberrant in a standard setting procedure (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 total score on a reading test and .44 with the total score on a mathematics test. Excluding judges with extreme C_i^* values had no effect on the recommended test standard.

Van der Flier (1982) used simulated data to investigate the usefulness of $ZU3$. In his first study, item score patterns were simulated on the basis of the item difficulties from two different populations (denoted populations I and II). $ZU3$ scores were determined on the basis of the π_g values in population I or II, and 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 on the basis of which a pattern was allocated to a population was unclear. Van der Flier (1982) found that approximately 70 % of the patterns were allocated to the correct population and that the percentage of correct allocations was not related to the total score.

Furthermore, the use of $ZU3$ was investigated in a cross-cultural setting. Kenyan and Tanzanian examinees were compared on the basis of a verbal reasoning test in Kiswahili. It was known that Kenyan examinees had less knowledge of Kiswahili than Tanzanian examinees. Van der Flier (1982) hypothesized that for examinees with low $ZU3$ scores (indicating aberrance), the test scores underestimated reasoning ability and that for groups of examinees with equal test scores, a more deviant group would obtain better results on a criterion variable. Van der Flier (1982) found that Kenyan people with high deviance scores on the verbal reasoning tests had better examination results (criterion) than would be expected on the basis of their verbal reasoning test scores (predictor). The additional information provided by the person-fit scores in predicting examination results, however, was rather modest.

Meijer (1994) used simulated data for comparing the detection rate of $U1$ and $U3$ and

found that the detection rates were comparable. Using simulated data, Meijer, Molenaar, and Sijtsma (1994) investigated the influence of test length, the type of aberrant responses, and the overall item discrimination on the detection rate of $U3$. They found that a priori defined aberrant item score patterns were easier to detect with longer tests and higher item discrimination. Moreover, the kind of aberrant behavior had a strong influence on the detection rate of $U3$. For example, nonfitting item score patterns were simulated by changing the 0 scores on the most difficult items into 1 scores (mimicking cheating) or by assigning a 1 score with a probability of .25 to each item (mimicking guessing). Cheaters were easier to detect than guessers.

Meijer (1996) used simulated data for investigating the influence of the type and the number of aberrant patterns in a calibration sample on the detection rate of $ZU3$. An increase in the number of aberrant simulees resulted in biased estimates of the $\pi_g s$ and in a decrease in the detection rate of $ZU3$. Furthermore, the type of misfit and the test length influenced the detection rate of $ZU3$. The use of an iterative procedure to re-estimate the proportion-correct score after removing aberrant patterns from the data was investigated. Item score patterns that were classified as aberrant were removed from the dataset and the proportion correct score was re-estimated until no clear improvement in the detection rate was found. Results suggested that this method can be used to improve the detection rate of $ZU3$ when aberrant examinees are present in a data.

Evaluation of Group-Based statistics

Group-based statistics classify a score pattern as aberrant when it is different from the other score patterns in a group. With the exception of $ZU3$, researchers chose the critical values for classifying a score pattern as aberrant by means of rules of thumb, based on the characteristics of the data. For example, Harnisch (1983) suggested that for C_i a value higher than .6 indicated aberrance. Harnisch and Linn (1983) labelled item score patterns with $C_i^* > .3$ as aberrant. These critical values, however, were based on one or two empirical data sets.

Criteria for the selection of useful statistics that were used by Harnisch and Linn (1981) and Rudner (1983) were (1) low correlation with the total score, and (2) detection rate. Harnisch and Linn (1981) concluded that of the statistics considered in their study, C_i^* was least related to the total score and was the most suitable statistic to detect aberrant

item score patterns. On the basis of the literature, a full comparison of the correlations between person-fit statistics and the total score, and of the rates of detection seems hardly possible. The studies are incomplete, and the characteristics of the datasets are not always clear.

The group-based statistics may be sensitive to nonfitting response behavior, but one drawback is that their null distributions are unknown (with the exception of $ZU3$) and, as a result, it cannot be decided on the basis of significance probabilities when a score pattern is unlikely given a nominal Type I error rate. In general, let t be the observed value of a person-fit statistic T. Then, the significance probability or probability of exceedance is defined as the probability under the sampling distribution that the value of the test statistic is smaller than the observed value: $p* = P(T \leqslant t)$ or larger than the observed value $p* = P(T \geq t)$ depending on whether low or high values of the statistic indicate aberrant response behavior. Although it may be argued that this is not a serious problem as long as one is only interested in the use of a person-fit statistic as a descriptive measure, a more serious problem is that the distribution of the numerical values of most group-based statistics is dependent on the total score (e.g., Drasgow, Levine, & McLaughlin, 1987). This dependence implies that when one critical value is used across total scores, the probability of classifying a score pattern as aberrant is a function of the total score, however, which obviously is undesirable.

To summarize, it can be concluded that the use of group-based statistics has been explorative, and with the increasing interest in IRT modeling person-fit increasingly has been investigated within the IRT context.

Person-Fit Measures Based on Item Response Theory

Statistics

Prerequisites

In IRT the probability of obtaining a correct answer on item g ($g = 1, ..., k$) is a function of the latent trait value (θ) and characteristics of the item such as the location δ (Hambleton & Swaminathan, 1985; van der Linden & Hambleton, 1997). This conditional probability $P_g(\theta)$ is the item response function (IRF). Further, we define the vector with item score

random variables $\mathbf{X} = (X_1, ..., X_k)$ and a realization $\mathbf{x} = (x_1, ..., x_k)$. IRT often assumes that the item scores are locally independent

$$P(\mathbf{X} = \mathbf{x}|\theta) = \prod_{g=1}^{k} P_g(\theta)^{x_g}[(1 - P_g(\theta)]^{1-x_g}. \quad (7)$$

For any cumulative probability distribution of θ, F(θ), θ can be integrated out, which yields

$$P(\mathbf{X} = \mathbf{x}) = \int_{\theta} \prod_{g=1}^{k} P_g(\theta)^{x_g}[(1 - P_g(\theta)]^{1-x_g} dF(\theta). \quad (8)$$

In order to have testable restrictions on the distribution of $\mathbf{X}$, specific choices for the $P_g(\theta)s$, for F(θ), or for both have to be made. Whereas F(θ) sometimes is chosen to be normal, $P_g(\theta)$ often is specified using the 1-, 2-, or 3-parameter logistic model (1-, 2-, 3-PLM). The 3-PLM is defined as

$$P_g(\theta) = \gamma_g + \frac{(1 - \gamma_g)\exp[\alpha_g(\theta - \delta_g)]}{1 + \exp[\alpha_g(\theta - \delta_g)]}, \quad (9)$$

where γ_g is the lower asymptote [γ_g is the probability of a 1 score for low-ability examinees (that is, $\theta \to -\infty$)]; α_g is the slope parameter (or item discrimination parameter); and δ_g is the item location parameter. The 2-PLM can be obtained by fixing $\gamma_g = 0$ for all items; and the 1-PLM or Rasch model can be obtained by additionally fixing $\alpha_g = 1$ for all items.

A major advantage of IRT models is that the goodness-of-fit of a model to empirical data can be investigated. Compared to group-based person-fit statistics, this provides the opportunity of evaluating the fit of item score patterns to an IRT model. To investigate the goodness-of-fit of item score patterns, several IRT-based person-fit statistics have been proposed.

Let $w_g(\theta)$ and $w_0(\theta)$ be suitable functions. Following Snijders (1998), a general form in which most person-fit statistics can be expressed is

$$\sum_{g=1}^{k} X_g w_g(\theta) - w_0(\theta). \quad (10)$$

To have a person-fit statistic with expectation 0, many person-fit statistics are expressed in the centered form

$$V = \sum_{g=1}^{k} [X_g - P_g(\theta)]\, w_g(\theta). \tag{11}$$

Note that, as a result of binary scoring, $X_g^2 = X_g$; thus, for a suitable function $v_g(\theta)$ statistics of the form

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

can be re-expressed as statistics of the form in Equation (10).

Residual-Based Statistics

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

$$v_g(\theta) = \frac{1}{kP_g(\theta)[1 - P_g(\theta)]}$$

results in

$$U = \sum_{g=1}^{k} \frac{[X_g - P_g(\theta)]^2}{kP_g(\theta)[1 - P_g(\theta)]}. \tag{12}$$

Note that the denominator 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. W is defined as

$$W = \frac{\sum_{g=1}^{k} [X_g - P_g(\theta)]^2}{\sum_{g=1}^{k} P_g(\theta)[1 - P_g(\theta)]}. \tag{13}$$

The difference between U and W is that Wright and Stone (1979) asssumed that W is less sensitive to the event of one unexpected response to an item with a difficulty far away from the ability of an examinee. Wright and Stone (1979) and Wright and Masters (1982) claimed that the transformation of U,

$$ZU = [\ln U + U + 1](df/8)^{-1}, \tag{14}$$

with $df = k - 1$, and the transformation of W,

$$ZW = 3(W^{1/3} - 1)/q + (q/3), \tag{15}$$

where q is the variance of W, are asymptotically standard normally distributed. The appropriateness of these transformations for approximating the normal distribution can be questioned, however, as will be discussed below.

Two related statistics were proposed by Smith (1985). Let a test be divided into A_s ($s = 1,..., S$) non-overlapping subsets of items, then the unweighted between-sets fit statistic is defined as

$$UB = \frac{1}{S-1}\sum_{s=1}^{S}\frac{\{\sum_{g\in A_s}[X_g - P_g(\theta)]\}^2}{\sum_{g\in A_s}P_g(\theta)\,[1 - P_g(\theta)]}. \tag{16}$$

Let m_s denote the number of items in subset A_s, then the unweighted within-sets fit statistic is defined as

$$UW = \frac{1}{m_s}\sum_{g\in A_s}\frac{[X_g - P_g(\theta)]^2}{kP_g(\theta)[1 - P_g(\theta)]}. \tag{17}$$

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

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

is asymptotically multivariate normal, and (2) the distribution of UB is asymptotically chi-square distributed with S degrees of freedom when θ is used and $S - 1$ degrees of freedom when the maximum likelihood estimate $\hat{\theta}$ is used. To investigate whether the asymptotic distributions hold reasonably well for tests of realistic length, empirical dis-

tributions were simulated for tests consisting of 40 items. Kogut (1988) concluded that the empirical distributions were accurate enough to approximate the asymptotic distributions. Interesting is that both UW and UB can be used as diagnostic tools investigating whether a priori specified subsets of items fit the IRT model (UW) or for testing the null hypothesis that a examinee's ability is the same across subgroups (UB).

Likelihood-Based Statistics

Most studies, to be discussed below, have been conducted using some suitable function of the log-likelihood function

$$l = \sum_{g=1}^{k} \{X_g \ln P_g(\theta) + (1 - X_g) \ln[1 - P_g(\theta)]\}. \tag{18}$$

This statistic, first proposed by Levine and Rubin (1979), was further developed and applied in a series of articles by Drasgow, Levine and colleagues (e.g., Drasgow, Levine, & Williams, 1985; Drasgow, Levine, & McLaughlin, 1991; Levine & Drasgow, 1982; Levine & Drasgow, 1983). Two problems exist when using l as a fit statistic. The first problem is that l is not standardized, implying that the classification of an item score pattern as normal or aberrant depends on θ. The second problem is that for classifying an item score pattern as aberrant, a distribution of the statistic under the null hypothesis of fitting response behavior, the null distribution is needed, and for l this distribution is unknown. Solutions proposed for these two problems are the following.

To overcome the problem of dependence on trait level and the problem of unknown sampling distribution, Drasgow et al. (1985) proposed a standardized version l_z of l which was less confounded with θ and which was purported to be asymptotically standard normally distributed; l_z was defined as

$$l_z = \frac{l - E(l)}{[Var(l)]^{\frac{1}{2}}}, \tag{19}$$

where $E(l)$ and $Var(l)$ denote the expectation and the variance of l, respectively. These quantities are given by

$$E(l) = \sum_{g=1}^{k} \{P_g(\theta) \ln [P_g(\theta)] + [1 - P_g(\theta)] \ln [1 - P_g(\theta)]\}, \tag{20}$$

and

$$Var(l) = \sum_{g=1}^{k} P_g(\theta)\,[1 - P_g(\theta)] \left[\ln \frac{P_g(\theta)}{1-P_g(\theta)}\right]^2. \tag{21}$$

Molenaar and Hoijtink (1990; 1996) argued that l_z is only standard normally distributed when the true θ values are used, but in practice θ is replaced by the maximum likelihood estimate $\hat{\theta}$. Using an estimate and not the true θ will have an effect on the distribution of a person fit statistic, as was shown by Molenaar and Hoijtink (1990), Nering (1995, 1997), and Reise (1995). These studies showed that when maximum likelihood estimates $\hat{\theta}$ were used, the variance of l_z was smaller than expected under the standard normal distribution using the true θ, particularly for tests up to moderate length (say, 50 items or less). As a result, the empirical Type I error was smaller than the nominal Type I error.

For the Rasch model, Molenaar and Hoijtink (1990, p. 96) showed that l_0 can be written as the sum of two terms. Given $\sum_{g=1}^{k} X_g = r$ (that is, given $\hat{\theta}$, which in the Rasch model only depends on the sufficient statistic r) one term is independent of the item score pattern and the other is dependent on it. Following their notation, the former part is denoted by d_0 and the latter by M, such that

$$l = d_0 + M$$

with

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

and

$$M = -\sum_{g=1}^{k} \delta_g X_g. \tag{22}$$

Note that if the distribution of l conditional on $\sum_{g=1}^{k} X = r$ is considered, d_0 is independent of $\mathbf{X}$, and l and M have the same ordering in $\mathbf{X}$. Because of its simplicity, Molenaar and Hoijtink (1990) used M rather than l as a person-fit statistic. Molenaar and Hoijtink (1990) proposed three approximations to the distribution of M: using (1) complete enumeration, (2) Monte Carlo simulation and, (3) a χ^2-distribution, where the mean, standard deviation, and skewness of M are taken into account; see Molenaar and

Hoijtink (1990) for the conditions when to use either one of these approaches. In line with this research, Liou and Chang (1992) proposed a network algorithm that enumerated all possible response patterns to construct exact tail probabilities for l and Bedrick (1997) derived alternative methods to approximate the first two moments of M.

Drasgow, Levine, and McLaughlin (1991) proposed a generalization l_{zm} of the l_z statistic for tests consisting of S unidimensional subtests or testlets. This statistic has a similar expression as l_z, but now the expectation and variance are taken over S subtests:

$$l_{zm} = \frac{\sum_{s=1}^{S}[(l^{(s)} - E(l^{(s)})]}{\sum_{s=1}^{S}[Var(l^{(s)})]^{1/2}}. \tag{23}$$

Although Drasgow et al. (1991) showed that l_{zm} was effective in detecting aberrant 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 testlets. In practical test situations, the use of l_{zm} suffers from the same problems as l_z : using $\hat{\theta}$ instead of θ will result in inappropriate approximations to probabilities of exceedance. Using l_z in the context of the 3-PLM, 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 that all have the form given in Equation (11) and for which the maximum likelihood estimate $\hat{\theta}$ was used instead of θ. It can easily be shown (Snijders, 1998) that $l_0 - E(l_0)$ can be written in the form of Equation (11) choosing

$$w_g = \left(\frac{P_g(\theta)}{1 - P_g(\theta)}\right).$$

Snijders (1998) derived expressions for the first two moments of the distribution: $E\left[V(\hat{\theta})\right]$ and $Var\left[V(\hat{\theta})\right]$; and performed a simulation study for relatively small tests consisting of 8 and 15 items, fitting the 2-PLM, and using maximum likelihood estimation for θ. The results showed that the approximation was satisfactory for $\alpha = 0.05$ and $\alpha = 0.10$, but that the empirical Type I error was higher than the nominal Type I error for smaller values of α.

Drasgow, Levine, & McLaughlin (1987) proposed two fit statistics that are sensitive

to the flatness of the likelihood function. The idea was that when there is no single value of θ that provides a good fit for an item score pattern, the likelihood function will be relatively flat. The first statistic (JK) is a normalized jackknife variance estimate. Let $\hat{\theta}$ denote the 3-PLM maximum likelihood estimate of θ, based on all k items in the test, and let $\hat{\theta}^*_{(g)}$ denote the estimate based on $k-1$ items remaining when item g is excluded. Then

$$\hat{\theta}^*_g \equiv k\hat{\theta} - (k-1)\hat{\theta}^*_{(g)} \text{ for } g = 1, \ldots, k.$$

and the jackknife estimate of θ is

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

with variance

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

Because there is more Fisher information about θ in some ranges of θ than in other ranges of θ, $Var(\hat{\theta}^*)$ depends on θ. Therefore, $Var(\hat{\theta}^*)$ was weighted by the Fisher information, $I(\hat{\theta})$, of which the reciprocal is the asymptotic variance of $\hat{\theta}$, which resulted in

$$JK = Var(\hat{\theta}^*)I(\hat{\theta}). \tag{24}$$

The second person-fit statistic was the ratio of the observed and the expected information:

$$O/E = \frac{\frac{\partial^2 l}{\partial \theta^2}|_{\theta=\hat{\theta}}}{I(\hat{\theta})}. \tag{25}$$

The idea behind this statistic was that if the likelihood l (see Equation 18) is flatter for aberrant responses than for normal responses, then the observed information is expected to be smaller than the expected information.

Drasgow et al. (1987) also proposed to use the variance of the mean number-correct

score of examinees who selected option d of item g (item-option variance: IOV):

$$IOV = Var(X_{dg}). \tag{26}$$

Conditional on θ, large values of IOV point at deviant behavior.

Statistics Based on the Caution Index

Tatsuoka and Linn(1983) derived several statistics which were similar to the caution index C_i discussed by Harnisch and Linn (1981) and which were adapted to IRT modeling. Let $\mathbf{X}_i$ be the vector of item scores of examinee i; let $\mathbf{X}_i^*$ be the theoretical Guttman vector, and let $\mathbf{n}$ be the vector with the item number-correct scores across examinees. The caution index can be written as

$$C_i = 1 - \frac{cov(\mathbf{X}_i, \mathbf{n})}{cov(\mathbf{X}_i^*, \mathbf{n})}. \tag{27}$$

Also, let $\mathbf{P}(\theta)$ be the vector with conditional probabilities $P_g(\theta)$ across items. A vector $\mathbf{P}(\theta)$ is defined for each θ. By norming against the covariance between the probability of a correct response under an IRT model and vector $\mathbf{n}$, statistic $ECI1$ was obtained as

$$ECI1 = 1 - \frac{Cov(\mathbf{X}_i, \mathbf{n})}{Cov[\mathbf{P}(\theta), \mathbf{n}]}. \tag{28}$$

Similarly $ECI2$ (and, likewise, $ECI3$) was obtained by taking the covariance (correlation) between an item score vector and the vector with the mean probability of correctly answering an item across n examinees, $\mathbf{G} = (G_1, ..., G_k)$ with elements $G_g = 1/n \sum_{i=1}^{n} P_g(\theta)$:

$$ECI2 = 1 - \frac{Cov[\mathbf{X}_i, \mathbf{G}]}{Cov[\mathbf{P}(\theta), \mathbf{G}]}, \tag{29}$$

and

$$ECI3 = 1 - \frac{Corr[\mathbf{X}_i, \mathbf{G}]}{Corr[\mathbf{P}(\theta), \mathbf{G}]}. \tag{30}$$

$ECI4$, $ECI5$, and $ECI6$ were obtained by taking the covariance or the correlation between the response vector X_i and $\mathbf{P}(\theta)$, resulting in the following statistics

$$ECI4 = 1 - \frac{Cov[\mathbf{X}_i, \mathbf{P}(\theta)]}{Cov[\mathbf{G}, \mathbf{P}(\theta)]}; \tag{31}$$

$$ECI5 = 1 - \frac{Corr[\mathbf{X}_i, \mathbf{P}(\theta)]}{Corr[\mathbf{G}, \mathbf{P}(\theta)]}; \tag{32}$$

and

$$ECI6 = 1 - \frac{Cov[\mathbf{X}_i, \mathbf{P}(\theta)]}{Var[\mathbf{P}(\theta)]}. \tag{33}$$

An important difference between these statistics is that $ECI2$ and $ECI3$ compare an individual item score pattern with the *mean* probability across persons and thus compare an individual item score pattern with group characteristics, whereas $ECI4$, $ECI5$, 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 and $ECI6$ is normed against the variance of $\mathbf{P}(\theta)$. $ECI3$ and $ECI5$ are similar to $ECI2$ and $ECI4$, with the difference that in $ECI3$ and $ECI5$ correlations are used instead of covariances. Tatsuoka (1984) derived the expectations and the variances of $ECI1$, $ECI2$, $ECI4$, and $ECI5$ and used these to obtain standardized versions of these indices (subtracting the expected values and dividing by the standard errors). These standardized indices were denoted $ECI1_z$, $ECI2_z$, $ECI4_z$, and $ECI5_z$.

Although it has sometimes be remarked that likelihood statistics and the ECI statistics are based on different approaches to person-fit (e.g., Harnisch & Tatsuoka, 1983; Kogut, 1986; Nering, 1997), it can be shown that both approaches are of the form of Equation (11). For example, the centered form of $ECI4$, that is, $ECI4 - E(ECI4)$ can be obtained by choosing

$$w_g = P_g(\theta) - \overline{P}(\theta),$$

where $\overline{P}(\theta) = 1/k \sum_{g=1}^{k} P_g(\theta)$, and the centered form of $ECI2$ can be obtained by choosing

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

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

Optimal Person-Fit Statistics

Levine and Drasgow (1988; see also Drasgow & Levine, 1986; Drasgow, Levine, & Zickar, 1996) proposed a method for the identification of aberrant item score patterns which was statistically optimal; that is, no other method can achieve a higher rate of detection at the same Type I error rate. A likelihood ratio statistic was determined which provided the most powerful test for the null hypothesis that an item score pattern is normal versus the alternative hypothesis that it is aberrant. The researcher in advance has to specify a model for normal behavior (e.g., the 1-,2-, or 3- PLM) and a model that specifies a particular type of aberrant behavior (e.g., a model in which violations of local independence are specified). The likelihood ratio statistic

$$\lambda(\mathbf{X}) = \frac{P(\mathbf{X} = \mathbf{x})_{aberrant}}{P(\mathbf{X} = \mathbf{x})_{normal}} \tag{34}$$

is calculated, and those patterns are classified as aberrant (1) which have the largest $\lambda(\mathbf{X})$ and (2) whose likelihoods under the model describing normal response behavior sum up to the α level.

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

$$P(\mathbf{X} = \mathbf{x} \mid \theta) = \mu(\theta) h(\mathbf{x}) \exp[\theta R(\mathbf{x})], \tag{35}$$

where

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

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

and $R(\mathbf{x})$ = number-correct score, Klauer (1995) modelled aberrant response behavior using the two-parameter exponential family, and introducing an extra person parameter η, as

$$P(\mathbf{X} = \mathbf{x} \mid \theta, \eta) = \mu(\theta, \eta)h(\mathbf{x}) \exp[\eta T(\mathbf{x}) + \theta R(\mathbf{x})], \tag{36}$$

where $T(\mathbf{x})$ depends on the particular alternative model considered. Using the exponential family of models, a uniformly most powerful test (e.g., 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 were θ_1 is an individual's ability on subtest A_1 and θ_2 is an individual's ability on subtest A_2. Under the Rasch model, it is expected that θ is invariant across subtests and thus H_0: $\eta = 0$ can be tested against H_1: $\eta \neq 0$. For this type of aberrant behavior $T(\mathbf{x})$ is the number-correct score on either one of the subtests.

Klauer (1995) also tested H_0 of equal item discrimination parameters for all persons against person-specific item discrimination and H_0 of locally independence against violations against local independence. Results showed that the power of these tests depended on the type and severeness of the violations. Violations against non-invariant ability (H_0: $\eta = 0$) were found to be the most difficult to detect. Liou (1993) discussed refinements using these types of tests.

Interesting in both the Levine and Drasgow (1988) and Klauer (1991, 1995) approaches is that model violations are specified in advance and that tests are proposed to investigate these model violations. This is different from the approach followed in most person-fit studies where the alternative hypothesis simply says that the null hypothesis is not true. 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-conform item score patterns against these alternative models. Another option is to first investigate which model violations are most detrimental to the use of the test envisaged and then test against the most serious violations (Klauer, 1995).

The Person-Response Function

Trabin and Weiss (1983; see also Weiss, 1973) proposed to use the person response function (PRF) to identify aberrant 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. In IRT, the item response function often 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 (1983) ordered items to increasing $\hat{\delta}$ values and then formed subtests of items by grouping items according to $\hat{\delta}$ values. For fixed $\hat{\theta}$, the observed PRF was constructed by determining, in each subtest, the mean probability of a correct response. The expected PRF was constructed by estimating according to the 3-PLM, in each subtest, the mean probability of a correct response. A large difference between the expected and observed PRFs was interpreted as an indication of nonfitting responses for that examinee.

Let k items be ordered by their δ values and let item rank numbers be assigned accordingly, such that

$$\delta_1 < \delta_2 < \ldots < \delta_k \tag{37}$$

Furthermore, let $\hat{\theta}$ be the maximum likelihood estimate of θ under the 3-PLM. Assume that A_s (s = 1, ..., S) ordered subtests can be formed, each containing m items; thus, $A_1 = \{1, \ldots, m\}$, $A_2 = \{m+1, \ldots, 2m\}, \ldots$, $A_S = \{k-m+1, \ldots, k\}$, and note that $S * m = k$. To construct the expected PRF, an estimate of the expected proportion of correct responses on the basis of the 3-PLM in each subtest is taken:

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

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

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

Within each subtest, for a particular $\hat{\theta}$ the difference of observed and expected correct scores is taken and this difference is divided by the number of items in the subtest. This

yields

$$D_s(\hat{\theta}) = m^{-1} \sum_{g \in A_s} \left[X_g - P_g(\hat{\theta}) \right], \text{ all } s = 1, ..., S. \tag{38}$$

Next, the Ds are added across subtests, which yields

$$D(\hat{\theta}) = \sum_{s=1}^{S} D_s(\hat{\theta}). \tag{39}$$

$D(\hat{\theta})$ was taken as a measure of an individual's fit to the model. For example, when an examinee copied the answers on the most difficult items, for such examinees scores on the most difficult subtests are likely to be substantially higher than suggested by the expected PRF. Related ideas were discussed by Lumsden (1977, 1978).

Klauer and Rettig (1990) expanded the methodology of Trabin & Weiss (1983) by proposing three person-fit statistics that were standardized and for long tests asymptotically followed a chi-square distribution. Let a test be divided into S subtests with $s = 1, ..., S$, and let the latent trait estimate for the total test (k items) be denoted by $\hat{\theta}$. One of the statistics was

$$\chi^2_{sc} = \sum_{s=1}^{S} \frac{V_s^2(\hat{\theta})}{I_s(\hat{\theta})}, \tag{40}$$

where $V_s(\hat{\theta})$ is of the form given in Equation (11),

$$V_s(\hat{\theta}) = \sum_{g \in A_S} \left[X_g - P_g(\hat{\theta}) \right] w_g(\hat{\theta}) \tag{41}$$

with

$$w_g(\theta) = \frac{dP_g(\theta)/d\theta}{P_g(\theta)\left[1 - P_g(\theta)\right]}, \tag{42}$$

and $I_s(\hat{\theta})$ is the estimated Fisher's information function. To test whether θ is invariant across subtests, the null hypothesis H_0: $\theta_1 = \theta_2 = ... = \theta_S$ is tested. Under H_0, χ^2_{sc} has a chi-square distribution with $df = S - 1$. Note that this test is similar to the method proposed by Trabin & Weiss (1983) but differs in that χ^2_{sc} is standardized and asymptotically chi-square distributed. Klauer and Rettig (1990) also proposed two related tests.

The first was the Wald test, which directly compares a person's ability estimates obtained from different subtests. The other was a likelihood ratio test. By means of Monte Carlo research Klauer and Rettig (1990) showed that the chi-square distribution of χ^2_{sc} was appropriate for tests of at least 80 items. For the Wald test and the likelihood ratio test the difference between the theoretical and empirical chi-square distributions was too large to be of practical use.

Research Using IRT-Based Person-Fit Statistics

Several studies have addressed the usefulness of IRT-based person-fit statistics. In most studies simulated data were used, and in some studies, empirical data were used. We will distinguish studies investigating the

1. detection rate of fit statistics and comparing fit statistics with respect to several criteria such as distributional characteristics and relation to the total score;
2. influence of item, test, and person characteristics on the detection rate;
3. relation between nonfitting score patterns and the validity of test scores; and
4. applicability of person-fit statistics to detect particular types of nonfitting item score patterns.

Although some studies may be categorized under more than one heading, we discuss it under the heading where it seems to have its largest contribution.

Studies Investigating Detection Rate and Comparing Fit-Statistics

Levine and Rubin (1979) evaluated statistic l in a study that simulated 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 normal examinees (generated using the 3-PLM) and changing these scores into 1 scores. Spuriously low-scoring examinees were simulated by sampling a fixed percentage of the item scores and rescoring the items as correct with a probability of 0.20. The percentages of items scores that were taken were 4, 10, 20, and 40. Levine and Rubin (1979) found that the larger the group of aberrant item scores the better l could distinguish normal from aberrant score patterns. They also found that spuriously high-scoring simulees were easier to detect than spuriously low-scoring simulees. This could be understood because more item scores were changed for the spuriously high-scoring

examinees.

In another study using l, Drasgow (1982) compared the detection rates of l using either the Rasch model or the 3-PLM to describe the data of the Graduate Record Examination. He found higher detection rates for examinees with spuriously low manipulated item scores than for examinees with spuriously high manipulated item scores. Furthermore, detection rates for this dataset were higher using the 3-PLM than using the Rasch model.

Harnisch and Tatsuoka (1983) used National Assessment of Educational Progress (NAEP) data on mathematics to investigate the correlations, scatterplots, and tests for curvilinearity for ECIa (a= $1, 2, ..., 5$), $ECI1_z$,$ECI2_z$,$ECI4_z$, U, W, l,and l_z. U was used under the 2-PLM and the 3-PLM and l was used under the 3-PLM and under the Normal-Ogive model (Hambleton & Swaminathan, 1985, pp. 35-36). Harnisch and Tatsuoka (1983) found that $ECI1$, $ECI2$, and $ECI4$ had standard deviations of approximately 1 and means of approximately .20. U correlated lowest with the other indices (approximately .10), and the other indices correlated between .50 and .98 with each other. l_z and l correlated highest with the total score: .36 and .27, respectively. Furthermore, they found the strongest curvilinear relationship between the total score and l and between the total score and W.

Drasgow, Levine, & McLaughlin (1987) used the 3-PLM for comparing the person-fit statistics l_z, ZU, ZW, C, IOV, JK, O/E, $ECI2_z$, and $ECI4_z$, with (1) optimal statistics, (2) their standardization (i.e., if the distribution of the statistics was comparable across θ) and (3) detection rate. To determine the detection rate, nonfitting item scores were simulated in a similar way as in the Levine and Rubin (1979) study. Detection rates were found by determining the proportions of aberrant item score patterns that were correctly identified as aberrant when various proportions of normal aberrant item score patterns were misclassified as aberrant. Drasgow et al. (1987) concluded that ZU, C, and IOV were poorly standardized compared to the other statistics. They also found that $ECI4_z$ was better standardized and had a higher detection rate than $ECI2_z$. Furthermore, it was found that the O/E and the JK statistics were reasonably well standardized, but that these statistics were quite ineffective for the detection of aberrant item score patterns. One of the most interesting results was that l_z, ZW, and $ECI4_z$ had high detection rates for spuriously high scoring examinees having low θ values and for low scoring examinees

having high θ values (e.g., the detection rate of $ECI4_z$ was .75 for low θ values at a Type I error of .01 and 30% spuriously high scoring examinees). However, these statistics were less sensitive to manipulated response patterns for θ values around the mean 0 (for the example given above the detection rate decreased from .75 to .51 for low θ values at a Type I error of .01). Optimal indices had detection rates from 50% to 200% higher than other indices for average-ability examinees and item score patterns with spuriously high or spuriously low test scores.

Rogers and Hattie (1987) investigated the detection rate of ZU and ZW. Transformations of both statistics were claimed (Wright & Stone, 1979) to be asymptotically standard normally distributed. Rogers and Hattie (1987) determined the detection rate of ZU and ZW using theoretical critical values for guessing, heterogeneity of the discrimination parameters, and multidimensionality. They concluded that ZW was insensitive to heterogeneity of the discrimination parameters and to multidimensionality and sensitive to guessing; ZU was insensitive to guessing, heterogeneity of the discrimination parameters and to multidimensionality. Detection rates increased by no more than 2% compared to normally responding examinees.

Noonan, Boss, and Gessaroli (1992) investigated the distributional characteristics and the empirical critical values of l_z, $ECI4_z$, and ZW as a function of the test length and the IRT model (2-PLM and 3-PLM). They found that both l_z and $ECI4_z$ had means and standard deviations (SD) that approximated the standard normal distribution. However, ZW had a mean over replications of approximately 1.00 but a SD between .144 and .232. Furthermore, they found that $ECI4_z$ and ZW were positively skewed and that l_z was negatively skewed, whereas the skewness of $ECI4_z$ was half the skewness of the other two statistics. They also found that for all three statistics, the critical values were affected by test length and IRT model. The critical values of ZW were most affected. They concluded that $ECI4_z$ best approximated the normal distribution and, moreover, was less affected by test length and the IRT model. l_z and $ECI4_z$ were highly correlated (.95), whereas $ECI4_z$ and ZW had the smallest correlation (.58). However, true θ values were used which makes the generalization to empirical distributions difficult.

Li and Olejnik (1997) compared the distribution of five person-fit statistics that were normally distributed assuming the Rasch model: l_z , $ECI2_z$, $ECI4_z$, ZU, and ZW. They found that (1) the statistics had low correlation with the total score; (2) the statistics were

positively skewed and deviated significantly from normality, where $ECI4_z$ was better normalized than $ECI2_z$; (3) l_z performed at least as well as the other statistics in detecting aberrant behavior; (4) examinees with spuriously low and spuriously high total scores were equally well detectable when unidimensional data were used, whereas detection rates of spuriously low total scores were lower than detection rates of spuriously high total scores when a multidimensional test was used; and (5) person-fit statistics were not very powerful in identifying aberrant item score patterns; l_z was most powerful and detected at most 67% of the aberrant item score patterns. However, because it was assumed that the true θ equaled the maximum likelihood estimate $\hat{\theta}$, these conclusions suffer from the same shortcomings as the earlier work by Drasgow, Levine, and colleagues. As was discussed above, when using the Rasch model as Li and Olejnik did, it is better to condition on the total score, which is independent of $\hat{\theta}$, and to use statistic M. This was done by Kogut (1987), who used simulated Rasch model data to show that the detection rate of statistic M for detecting aberrant item score patterns was higher than l_z.

Trabin & Weiss (1983) applied the PRF approach to a 216 item vocabulary test which had been administered to 151 graduate students. To investigate whether the responses were in agreement with the 3-PLM, they used $D(\theta)$ for evaluating for each student the discrepancy between the observed and the expected PRF and assumed that $D(\theta)$ was chi-squared distributed. Some students had significant chi-squares but the cause of aberrance could not be explained.

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

Influence of Item, Person, and Test Characteristics

Several simulation studies investigated the detection rate and the distributional characteristics of person-fit statistics as a function of person and test characteristics. In general, the detection rate was defined as the percentage of aberrant simulees classified as aberrant by a particular statistic.

Levine and Drasgow (1982, 1983) investigated if (1) the detection rate of l was influenced by using estimated item parameters instead of using true parameters, and if (2) the presence of aberrant item score patterns influenced the item parameter estimates and the detection rate. Response vectors were simulated according to the 3-PLM using the estimated item parameters from a previous calibration study of the SAT (Verbal). Aberrant item score vectors were simulated by randomly selecting from each vector 20% of the item scores (0s and 1s) and changing these item scores with a probability of .20 (1s became 0s and 0s became 1s). They concluded that the detection rate of l was not seriously affected by the estimated item parameters and by the presence of nonfitting item score patterns. Kogut (1987), however, concluded from his simulation study that, as a result of the presence of deviant item score patterns in the sample, the power of l_z and M was seriously reduced. Possible explanation for the different results of both studies were the different statistics that were used and the different numbers of simulated item score patterns in both studies. In the Levine and Drasgow study, the percentage of nonfitting response vectors was 6.7, and in the Kogut study this percentage was 20. The higher percentage of nonfitting item score patterns may have reduced the power. Furthermore, the type of nonfitting item score vectors also may have been responsible for reduced power.

Reise and Due (1991) found that longer tests and larger spread between the item difficulties resulted in higher detection rates for l_z. They simulated item scores with less Fisher information for estimating θ than predicted by the parameters of an IRT model; that is, item scores were simulated using different levels of the α-parameter (which is related to item information, Hambleton & Swaminathan, 1985, p. 105). Furthermore, they varied test length from 7, 21, 35, to 49 items and also varied the spread in the δs and the γs. Reise and Due (1991) concluded that test length, spread of the δs and the value of the γs each affected the detection rate of l_z. They found that, in general, longer tests, larger spread of the δs and low γs values resulted in higher detection rates. Furthermore, they concluded that l_z obtained its lowest detection rate for low θ values.

Parsons (1983) investigated the effectiveness of a transformed version of l to detect simulated aberrant item score patterns on a personality inventory, the Job Descriptive Index, that measured satisfaction with multiple facets of a job. Data were generated according to the 2-PLM using the estimated item parameters from an empirical calibration sample. Twenty out of the 60 items were selected and for these items, scores were gen-

erated with a probability of .30 of obtaining the correct response. Results indicated that higher detection rates were obtained at higher θs. This could be explained because for these simulees more item scores were changed. Furthermore, it was found that the variance of the total score for aberrant item score patterns was lower than for normal patterns. The explanation was that aberrant item scores are probably uncorrelated with each other and this reduced the variance of the total score compared to the total score on a set of items that are correlated.

Smith (1985) compared robust estimators with the person-fit statistics. Robust estimators correct for unexpected responses and weigh these unexpected responses less to obtain a representative latent trait estimate. Smith concluded that it is better to use person-fit analysis because the robust estimators introduce a bias in the estimation of the latent trait.

Reise (1995) investigated the detection rate of the l_z person-fit statistic as a function of using true θ and several estimates of θ: maximum likelihood estimation (MLE), expected a posteriori (EAP) estimation and biweight (BIW) estimation. To estimate θ, datasets were simulated based on the estimated item parameters of four personality scales that fit the 2-PLM. Reise found that using true θ resulted consistently in the highest detection rate for l_z. The detection rate of l_z differed between the three estimation methods, but this difference depended on the type of test, the θ level, and the percentage of nonfitting responses. Reise also found that BIW estimation typically resulted in a somewhat higher detection rate than EAP and MLE. Meijer and 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 aberrant behavior. They found that the presence of aberrant item score patterns influenced the bias in $\hat{\theta}$, and this depended heavily on the type of misfit and the θ level. It was also found that 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 nine IRT-based statistics, which were $ECI1$, $ECI2$, $ECI4$, their standardized versions, and l, l_z, and U in distinguishing among the following types of empirical item score patterns: item scores of an uncoop-

erative group, item scores of a cooperative group and item scores of a group in which scores had randomly been generated item scores. The 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 only administered for research and development purposes. Except for statistic U, Birenbaum (1985) found significant differences in the mean value of the other statistics between the three groups. The correlation between the standardized indices was high (.90). However, l and U had a low correlation of .10. Most statistics had low correlations with the total score (between .13 and .22). Curvilinearity between the person-fit statistics and the total score was rejected for none of the three unstandardized ECIs. Largest curvilinearity was detected for l, indicating that this index yielded the most inflated values at both extremes of the ability scale.

Birenbaum (1986) investigated the relation between four person-fit statistics, $ECI1_z$, $ECI2_z$, $ECI4_z$, and l_z on the one hand, and the scores on an anxiety scale and a lie scale of the MMPI, and a general ability test on the other hand. It was hypothesized that a sample of examinees with low anxiety scores but with high lie scores has a less appropriate item score pattern on an ability test than low-anxiety examinees who scored low on a lie scale, because persons with high lie scores have the desire to deliberately impress the assessor by saying that they have low anxiety, but they cannot conceal the effect of their anxiety on the cognitive reasoning test scores. Birenbaum (1986) found high correlations between the four person-fit statistics (between .97 and .99). Furthermore, low correlations were found between the scores on the lie scale and the fit statistics (.10) and between the scores on the anxiety scale and the fit statistics (.14). Scores on the ability scale correlated .50 with the fit indices. There was indeed a significant difference between the mean scores of the person-fit statistics between the two groups, where examinees with low anxiety and high lie scores were found to be more aberrant than examinees with low anxiety and high lie scores.

Hoijtink (1987) investigated the effect of nonfitting item score patterns on the item fit to the Rasch model. The item score patterns were from two empirical datasets from a questionnaire measuring neurological and ophthalmic skills for general practitioners. Aberrant item score patterns were removed from the dataset and it was investigated whether this resulted in a better fit of nonfitting *items* to the Rasch model. To minimize the dan-

ger of adapting the data to the model, item score patterns only were removed under the condition that they should be classified as aberrant both under the original and improved item estimates and under the condition that the fit of the dataset as a whole improved after removing nonfitting examinees. Hoijtink showed that removing nonfitting item score patterns resulted for some items in a better fit to the model. However, it could not be explained why some examinees answered the questionnaires in a deviant way, as was done in the Birenbaum (1985) study.

Rudner, Bracey, and Skaggs (1996) investigated the use of statistic W in the context of 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 and Waller (1993) explored the use of l_z in personality measurement by analyzing empirical data of the Multidimensional Personality Questionnaire (Tellegen, 1982). Three possible applications of person-fit measurement in the context of personality research were discussed: detection of measurement error, detection of variation due to faulty responding, and detection of variation due to inappropriateness of the personality trait measured by the test for describing several examinees. Reise and Waller (1993) noted that it is difficult to distinguish persons not fitting the particular trait from misfit due to error of measurement or faultiness. To reduce the chances that misfit of a person's item score pattern was attributed to measurement error or faulty responding, they used unidimensional subscales and information from detection scales that flag persons exhibiting inconsistent answering behavior. By means of l_z persons could be identified not responding according to the 2-PLM and who were not flagged by inconsistency scales. However, the accuracy of the classification could not be evaluated because it was unknown which persons really behaved in an aberrant way.

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

Molenaar and Hoijtink (1996) investigated the use of statistic M given in Equation (22) in the context of the Rasch model for a test in which four- to seven-year-olds had

to indicate which of the three pictures presented was consistent with the item. Each picture consisted of a number of balls and stars which were colored white and black. They simply identified patterns with low probability of exceedance. For example, ordering the items from easy to difficult they identified on an 11-item test the response pattern (00000010011), which had a significance probability of .002, and concluded that this pattern was a candidate for closer inspection.

Validity of Test Scores and Aberrant Response Behavior

The effect of deviant response behavior on validity and decision-making was investigated in several studies. The importance of the relation between deviant response behavior and decision making was underlined by Drasgow and Guertler (1987). They argued that over- or underestimating $\hat{\theta}$ may have serious consequences. Overestimating $\hat{\theta}$ may result in selecting persons that are not able to fulfill a job and underestimating $\hat{\theta}$ may be expensive for the company due to extra selection efforts that are 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 normal and aberrant item score patterns. On the basis of the probabilities of score patterns under these distributions, the utility could be estimated and the critical value of a statistic could be determined in line with the estimated utility.

Schmitt, Cortina, and Whitney (1993) investigated whether aberrant item score patterns may distort both estimates of criterion-related validity and estimates of the relationship between trait levels and performance constructs. Using l_z and the 3-PLM, and four empirical datasets, they found little or no improvement of the correlation between the predictor and the criterion when aberrant 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 (normal or aberrant), and (3) their cross products, showed for some data sets a significant interaction term, implying that l_z scores may improve prediction. Meijer (1997) used simulated data for investigating the relation between aberrant response behavior and validity of test scores. He concluded that nonfitting 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 nonfitting item score patterns is relatively high (at

least .15 or higher). However, using l_z for removing aberrant item score patterns from a predictor test appeared to have little impact on the validity coefficient with the criterion test. 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 aberrant item score patterns remained in the sample. Meijer (1998) used $ZU3$ to identify persons with unexpected item scores on empirical selection data. It was shown that, in general, persons with inconsistent item scores are less predictability than persons with consistent item scores. Both persons with lower criterion scores and persons with higher criterion scores than predicted could be identified.

Statistics for Detecting Answer Copying

Person-fit statistics can be used for identifying individuals with item score patterns that are unlikely given the IRT model under consideration. Levine & Rubin (1979), Hulin et al. (1983) have suggested that these statistics also can be used to detect answer copying. However, to detect answer copying also methods are available that were designed especially for this purpose. We will not extensively discuss answer copying statistics because this was already excellently done by Frary (1993).We only mention two examples. Most answer copying statistics directly compare the item score patterns of two examinees and determine the number of item scores they have in common. Large proportions of similar item scores suggest answer copying.

One of the most promising statistics in this research area is the g_2 statistic (Frary, Tideman, & Watts, 1977). To calculate this statistic, a copier (i) and a source (j) have to be specified in advance. Let N_{ij} denote the number of items that is answered identically by persons i and j, and let $X_{gi} = x_{gi}$ and $X_{gj} = x_{gj}$ denote the realization of an item score of person i and an item score of person j, respectively. Then $N_{ij} = \sum_{g=1}^{k} T(X_{gi} = X_{gj})$, where $T = 1$ if i and j select the same alternative to item g, and $T = 0$ otherwise. Furthermore, let $\mathbf{X}_j$ denote the score pattern of person j. Treating $\mathbf{X}_j$ as fixed, g_2 is given by

$$g_2 = \frac{N_{ij} - E(N_{ij}|\mathbf{X_j})}{Var(N_{ij}|\mathbf{X_j})^{\frac{1}{2}}}, \tag{43}$$

with

$$E(N_{ij}|\mathbf{X_j}) = \sum_{g=1}^{k} P_i(X_{gi} = X_{gj}|\mathbf{X}_j) \tag{44}$$

and

$$Var(N_{ij}|\mathbf{X_j}) = \sum_{g=1}^{k} P_i(X_{gi} = X_{gj}|\mathbf{X}_j)\,[1 - P_i(X_{gi} = X_{gj}|\mathbf{X}_j)]\,. \tag{45}$$

Note that in determining $E(N_{ij}|\mathbf{X_j})$ and $Var(N_{ij}|\mathbf{X}_j)$ it is assumed that local independence holds given $\mathbf{X}_j$. Also note that a latent trait is not assumed. The obvious problem is here how to determine $P_i(X_{gi} = X_{gj}|\mathbf{X}_j)$. Frary et al. (1977) estimated these probabilities using estimates of π_g and the distractor answering proportions, and the ratio of the copier's number-correct score to the mean number-correct score for all examinees. As Wollack (1997) pointed out, g_2 does not take the trait level of the copier into account and, as a result, g_2 implicitly assumes that the item discriminations and the probabilities of selecting a particular distractor are constant across examinees with different trait levels.

As an alternative, Wollack (1997) proposed a statistic, ω, similar to g_2, but the probabilities associated with each response were determined using the nominal response model (Bock, 1972). This model specifies the probability of an examinee i with trait level θ_i selecting alternative u of item g. Treating $\mathbf{X}_j$ as fixed,

$$\omega = \frac{N_{ij} - E(N_{ij}|\theta_i, \mathbf{X_j})}{Var(N_{ij}|\theta_i, \mathbf{X_j})^{\frac{1}{2}}}, \tag{46}$$

where

$$E(N_{ij}|\theta_i, \mathbf{X_j}) = \sum_{g=1}^{k} P(X_{gi} = X_{gj}|\theta_i, \mathbf{X_j}) \tag{47}$$

and

$$Var(N_{ij}|\theta_i, \mathbf{X_j}) = \sum_{g=1}^{k} P(X_{gi} = X_{gj}|\theta_i, \mathbf{X_j})[1 - P(X_{gi} = X_{gj}|\theta_i, \mathbf{X_j})]. \tag{48}$$

ω is assumed to be asymptotically normally distributed. Using simulated data, Wollack (1997) found that the empirical Type I error rate was lower than the nominal Type I error rate. For most cases, however, this empirical Type I error rate was more in agreement with the nominal Type I error rate than similar results for the empirical and nominal Type

I error rates for g_2. As a result, the detection rate of ω also was higher than of g_2.

Because g_2 and ω were defined so as to detect specific types of aberrant behavior, for this purpose they are more powerful than general person-fit statistics. Some problems, however, seem to justify more research in this area. For example, as Wollack (1997) noted, when copying is involved, θ_i is confounded with θ_j and, because θ_i is taken as the copier's trait level to determine the probabilities used in ω, this statistic also may be confounded. In fact, the copier completes the test using two different θ values, his/her own and that from the source. Before answer copying is investigated, it may therefore be interesting to test by means of a person-fit statistic how serious the unidimensionality assumption is violated; that is, the fit of a person to the model should be investigated first. When an item score pattern fits the model, the researcher can have confidence in the validity of θ and its use in ω; when an item score pattern does not fit the model, care should be taken in interpreting ω, because the interpretation of θ may be ambiguous. Furthermore, because the normality assumption is based on asymptotic sampling theory, ω seems to be less suited for short tests (say, 40 items or less). Moreover, because the standardization of ω is similar to the standardization of the person-fit statistic l_z, using $\hat{\theta}$ instead of θ also is likely to influence the distribution of ω.

Discussion

Which statistic should be used ?

We discussed several methods that can be used to investigate the aberrance of individual item score patterns under particular IRT models. The methods ranged from statistics for testing whether an item score pattern is in agreement with the other patterns in the sample, to methods for investigating whether persons have been copying the correct answers from other examinees. Depending on the type of data and the problems envisaged, a researcher may choose a particular statistic, although not all statistics have equally favorable properties in a statistical sense. For example, for short tests and tests of moderate length (say, 10-60 items) and using the standard normal distribution, due to the use of $\hat{\theta}$ rather than θ for most statistics the nominal Type I error rate is not in agreement with the empirical Type I error rate. Recently, Snijders (1998) proposed statistical theory for correcting the bias caused by using the maximum likelihood estimate $\hat{\theta}$ rather than θ. In general,

sound statistical methods have been derived for the Rasch model, but because this model is rather restrictive to empirical data, the use of these statistics also is restricted.

In general, it may be wise to first investigate possible threats to the fit of individual item score patterns before using a particular person-fit statistic. If one suspects that answer copying is a realistic threat, one of the answer copying statistics can be used as an alternative to a person-fit statistic. As another example, if violations against local independence are expected, one of the methods proposed by Klauer (1991) may be used instead of a general statistic such as proposed by Molenaar and Hoijtink (1990). Not only are tests against a specific alternative more powerful than general statistics, also the type of deviance is easier to interpret. Statistics like the M statistic proposed by Molenaar and Hoijtink (1991) are helpful in situations when the researcher has no idea which threats are most important. Statistics like UB (Equation 16) and UW (Equation 17) or the person-response function can be used as diagnostic tools to test whether item score patterns on a priori specified subtests fit the IRT model.

A drawback of some person-fit statistics is that only deviations against the model are tested. This may result in interpretation problems. For example, item score patterns not fitting the Rasch model may be described more appropriately by means of the 3-PLM or may be flagged by a statistic for answer copying. If the Rasch model does not fit the data, other explanations are possible. Because in practice it is often difficult, if not impossible, to substantially distinguish different types of item score patterns and/or to obtain additional information using background variables, a more fruitful strategy may be to test against specific alternatives.

Almost all statistics are of the form given in Equation (11) but the weights are different. The question then is which statistic should be used? The literature told us that the use of a statistic depends on what kind of model is used. Using the Rasch model, the theory presented by Molenaar and Hoijtink and their statistic M are a good choice. Statistic M should be preferred over statistics l_z or ZW because the critical values for M are more accurate than those of l_z and ZW or those of other statistics. Statistic M is available in the computer program RSP (Glas & Ellis, 1993) so that the practitioner can easily add the person-fit values to his/her dataset. With respect to the 2-PLM and 3-PLM, all statistics proposed suffer from the problem that the standard normal distribution is inaccurate when $\hat{\theta}$ is used instead of θ. This seriously reduces the applicability of these statistics. The the-

ory recently proposed by Snijders (1998) may help the practitioner to obtain the correct critical values. Another argument for the use of a likelihood-based statistic is that it is an increasing function of the probability of a score pattern under a model. It can easily be shown that residual-based statistics like the ones given in Equations (16) and (17) do not reflect the probability 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, $ZU3$ may be preferred over the other fit statistics (like C^*) because this statistic is also an increasing function of the probability of the score pattern and, moreover, the distribution of $ZU3$ is known to be standard normal conditional on the total score. However, it is unknown whether the empirical distribution is in agreement with the theoretical distribution when nonparametric IRT models are used.

Can Person-Fit Statistics Improve Measurement Practice ?

The aim 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. The first requirement thus is that person-fit statistics are sensitive to nonfitting item score patterns. After having reviewed the studies using simulated data, it can be concluded that detection rates are highly dependent on (1) the type of aberrant response behavior, (2) the θ value, and (3) the test length. When item score patterns do not fit an IRT model, high detection rates can be obtained in particular for extreme θs, even when Type I errors are low (e.g., .001). The reason is that for extreme θs deviations from the expected item score patterns tend to be larger than for moderate θs. As a result of this pattern misfit, the bias in $\hat{\theta}$ tends to be larger for extreme θs than for moderate θs (Meijer & Nering, 1997). The general finding that detection rates for moderate θs tend to be lower than for extreme θs thus is not such a bad result and certainly puts the disappointment some authors (e.g., Reise, 1995) expressed about low detection rates for moderate θs in perspective.

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 test takers lacking motivation, may produce deviant item score patterns that are unlikely given the model. However, again it depends on the degree of aberrance of response behavior how useful person-fit statistics really are. We agree with some authors (Rudner et al., 1996; Reise & Flannery, 1996)

that new empirical research is needed, but it should be noted that more empirical studies do *not* provide the answer to the question whether person-fit statistics can be helpful in improving measurement practice. Empirical studies can illustrate the use of a person-fit statistic. For example, an empirical study may show that examinees that are unmotivated to fill out a questionnaire can be detected using a particular person-fit statistic. Whether person-fit statistics can help the 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 aberrant. (1) Instead of reporting one ability estimate for an examinee, several ability estimates can be reported on the basis of subtests that are in agreement with the model; (2) modify the item score pattern (for example eliminate the unreached items at the end) and re-estimate ability; (3) do not report the ability estimate and retest a person; or (4) decide that the error is small enough for the impact on the ability to be marginal. This decision can be based on comparing the error introduced by measurement disturbance and the standard error associated with each ability estimate. Which of these actions is taken very much 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

When reviewing the person-fit literature some suggestions for future research come to mind:

1. With respect to the distributional characteristics of the person-fit statistics, for the 1-PLM sound statistical theory helps the researcher to decide when an item score pattern is improbable. For the 2-PLM and the 3-PLM problems exist because $\hat{\theta}$ instead of θ is used and this makes the estimation of the probability of exceedance unreliable. Research of methods is needed that correct for using $\hat{\theta}$. Snijders (1998) proposed a correction of the variance of a group of person-fit statistics. However, also the skewness and kurtosis should be taken into account, especially when the nominal Type I levels are small. In that case, nominal Type I error levels and empirical Type I error levels are not in agreement (e.g. van Krimpen-Stoop & Meijer, in press-a).
2. With the increasing use of computerized adaptive testing (CAT) research, the use of

person-fit statistics in this context also may be investigated. McLeod & Lewis (1998) showed that it indeed makes a difference when a person has preknowledge of some of the items in a CAT and, as a result of that, can obtain a higher total score. Therefore, detection of such persons is important. Nering (1997) and van Krimpen-Stoop and Meijer (in press-a) showed that in a CAT the distributional characteristics of existing person-fit statistics are far off the expected distributions. Moreover, the characteristics of a CAT are unfavorable for person-fit research: there are relatively few items compared to a paper-and-pencil test and this results in a lower detection rate. Furthermore, for each examinee the spread in the item difficulties is small by definition. McLeod & Lewis (1998) discussed a Bayesian approach for the detection of examinees with preknowledge of the items. More research, however, is needed. A possibility is to work with statistics that are especially designed for a CAT. For example, in a CAT it is assumed that there is an alternation of correct and incorrect responses. A number of consecutive correct or incorrect answers is unexpected and may be the result of aberrant response behavior. Person-fit statistics that are especially designed may be more powerful than "conventional" person-fit statistics, and the statistical properties of the former statistics should be less susceptible to the characteristics of a CAT (van Krimpen-Stoop & Meijer, in press-b).

3. Few studies analyze empirical data using person-fit statistics. An explanation may be that person-fit statistics only inform the researcher that a score pattern does not fit the model without giving extra information about the type of aberrance. There is also the difficulty of distinguishing between examinees with item score patterns for whom the wrong IRT model is used and those whose item score patterns can be explained using additional information. Studies are needed that analyze empirical data together with background variables to obtain extra information about the type of aberrance. Reise and Flannery (1996) mentioned the application of person-fit research in cross-cultural studies to investigate the scalability of examinees with a different ethnic background.
4. In the context of nonparametric IRT modeling, few statistics exist that test a response pattern against model assumptions using known statistical properties. Much more research is needed here to obtain sound statistical methods.
5. One of the biggest problems of using person-fit statistics in practice is the relatively low power of these statistics in detecting aberrant item score patterns. Testing against

a specified alternative may be a solution. More information is needed, however, about the influence of aberrant response behavior on the total score on a test.

6. To enhance the interpretation of nonfitting item score patterns it may be possible to determine 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. Aberrant response patterns may be more easily interpreted using such external frames of reference.
7. The person response function also may be used to enhance the interpretation of aberrant item score patterns. Because a plot of the observed and expected response functions immediately clarifies which groups of observed responses disagree with the expected responses, the researcher may more easily hypothesize the explanation of the aberrant item score patterns. More research is needed here.

Acknowledgment

The authors express their gratitude to Edith M.L.A. van Krimpen-Stoop for her comments on an earlier draft of this paper.

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.

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

Cliff, N. (1983). Evaluating Guttman scales: Some old and new thoughts. In H.

Wainer & S. Messick: *Principals of modern psychological Measurement: A Festschrift for Frederic M. Lord* (pp. 283-300). Hillsdale N.J.: 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 for some multidimensional test batteries. *Applied Psychological Measurement,* 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.

Frary, R. B. (1993). Statistical detection of multiple-choice answer copying answer copying: Review and commentary. *Applied Measurement in Education, 6,* 153-165.

Frary, R. B., Tideman, T. N., & Watts, T. M. (1977). Indices of cheating on multiple choice tests. *Journal of Educational Statistics, 2,* 235-256.

Glas, C.A.W., & Ellis, J.L..(1993). *User's manual RSP.* Groningen, The Netherlands; iec 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: 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. In R. K. Hambleton: *Applications of item response theory*. Vancouver: Kluwer.

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

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.

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. Agreement coefficients as indices of dependability for domain-referenced tests: *Applied Psychological Measurement*, *4*, 105-126.

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.

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. Mole-

naar. *Rasch models, foundations, recent developments, and applications*, 97-110. New York: Springer-Verlag.

Kogut J. (1986). Review of IRT-based indices for detecting and diagnosing aberrant response patterns (Research Report 86-4). Enschede: University of Twente, department of Education.

Kogut, J. (1987). Detecting aberrant response patterns in the Rasch model. (Research Report 87-3). Enschede: University of Twente, Department of Education.

Kogut, J. (1988). Asymptotic distribution of a person-fit statistic. (Report 88-13). Enschede: 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 & 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). *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.

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

Meijer, R. R., Molenaar, I. W., & Sijtsma, K. (1994). Item, test, person and group characteristics and their influence on nonparametric appropriateness measurement. *Applied Psychological Measurement, 18,* 111-120.

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

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., & Meijer, R. R. (1998). A comparison of the person response function and the l_z person-fit statistic. *Applied Psychological Measurement, 22,* 53-69

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

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*. NY: McGraw-Hill.

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

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

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

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

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 assessement 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 (1978). *The construction and interpretation of S-P tables*. Tokyo: Meiji Tosho.

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

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.

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.

Snijders, T. (1998). *Asymptotic distribution of person-fit statistics with estimated person parameter.* Unpublished report, University of Groningen, The Netherlands.

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. In D.J. Weiss (Ed.), *New horizons in testing*. New York: Academic Press.

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

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (in press-b). Detecting person-misfit in adaptive testing using statistical process control techniques. In: W.J. van der Linden and C.A.W. Glas, *Computerized Adaptive Testing: theory and practice*. Boston: Kluwer Academic Publishers.

van der Flier, H. (1977). Environmental factors and deviant response patterns. In Y P. Poortinga (Ed.). *Basic Problems in Cross-cultural Psychology*. Amsterdam: Swets & Zeitlinger.

van der Flier, H. (1980). *Vergelijkbaarheid van individuele testprestaties* [*Compara-*

bility of individual test performance]. Lisse: Swets & 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*. NY: Springer Verlag.

Weiss, D. J. (1973). *The stratified adaptive computerized ability test* (research report 73-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometeric Methods Program, September 1973.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*, 97-116.

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

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

Wollack (1997). A nominal response model approach for detecting answer copying. *Applied Psychological Measurement, 21*, 307–320.

Titles of Recent Research Reports from the Department of Educational Measurement and Data Analysis. University of Twente, Enschede, The Netherlands.

RR-99-01	R.R. Meijer & K. Sijtsma, *A Review of Methods for Evaluating the Fit of Item Score Patterns on a Test*
RR-98-16	J.P. Fox & C.A.W. Glas, *Multi-level IRT with Measurement Error in the Predictor Variables*
RR-98-15	C.A.W. Glas & H.J. Vos, *Adaptive Mastery Testing Using the Rasch Model and Bayesian Sequential Decision Theory*
RR-98-14	A.A. Béguin & C.A.W. Glas, *MCMC Estimation of Multidimensional IRT Models*
RR-98-13	E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *Person Fit based on Statistical Process Control in an AdaptiveTesting Environment*
RR-98-12	W.J. van der Linden, *Optimal Assembly of Tests with Item Sets*
RR-98-11	W.J. van der Linden, B.P. Veldkamp & L.M. Reese, *An Integer Programming Approach to Item Pool Design*
RR-98-10	W.J. van der Linden, *A Discussion of Some Methodological Issues in International Assessments*
RR-98-09	B.P. Veldkamp, *Multiple Objective Test Assembly Problems*
RR-98-08	B.P. Veldkamp, *Multidimensional Test Assembly Based on Lagrangian Relaxation Techniques*
RR-98-07	W.J. van der Linden & C.A.W. Glas, *Capitalization on Item Calibration Error in Adaptive Testing*
RR-98-06	W.J. van der Linden, D.J. Scrams & D.L.Schnipke, *Using Response-Time Constraints in Item Selection to Control for Differential Speededness in Computerized Adaptive Testing*
RR-98-05	W.J. van der Linden, *Optimal Assembly of Educational and Psychological Tests, with a Bibliography*
RR-98-04	C.A.W. Glas, *Modification Indices for the 2-PL and the Nominal Response Model*
RR-98-03	C.A.W. Glas, *Quality Control of On-line Calibration in Computerized Assessment*
RR-98-02	R.R. Meijer & E.M.L.A. van Krimpen-Stoop, *Simulating the Null Distribution of Person-Fit Statistics for Conventional and Adaptive Tests*
RR-98-01	C.A.W. Glas, R.R. Meijer, E.M.L.A. van Krimpen-Stoop, *Statistical Tests for Person Misfit in Computerized Adaptive Testing*

RR-97-07 H.J. Vos, *A Minimax Sequential Procedure in the Context of Computerized Adaptive Mastery Testing*

RR-97-06 H.J. Vos, *Applications of Bayesian Decision Theory to Sequential Mastery Testing*

RR-97-05 W.J. van der Linden & Richard M. Luecht, *Observed-Score Equating as a Test Assembly Problem*

RR-97-04 W.J. van der Linden & J.J. Adema, *Simultaneous Assembly of Multiple Test Forms*

RR-97-03 W.J. van der Linden, *Multidimensional Adaptive Testing with a Minimum Error-Variance Criterion*

RR-97-02 W.J. van der Linden, *A Procedure for Empirical Initialization of Adaptive Testing Algorithms*

RR-97-01 W.J. van der Linden & Lynda M. Reese, *A Model for Optimal Constrained Adaptive Testing*

RR-96-04 C.A.W. Glas & A.A. Béguin, *Appropriateness of IRT Observed Score Equating*

RR-96-03 C.A.W. Glas, *Testing the Generalized Partial Credit Model*

RR-96-02 C.A.W. Glas, *Detection of Differential Item Functioning using Lagrange Multiplier Tests*

RR-96-01 W.J. van der Linden, *Bayesian Item Selection Criteria for Adaptive Testing*

RR-95-03 W.J. van der Linden, *Assembling Tests for the Measurement of Multiple Abilities*

RR-95-02 W.J. van der Linden, *Stochastic Order in Dichotomous Item Response Models for Fixed Tests, Adaptive Tests, or Multiple Abilities*

RR-95-01 W.J. van der Linden, *Some decision theory for course placement*

RR-94-17 H.J. Vos, *A compensatory model for simultaneously setting cutting scores for selection-placement-mastery decisions*

RR-94-16 H.J. Vos, *Applications of Bayesian decision theory to intelligent tutoring systems*

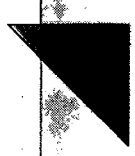
RR-94-15 H.J. Vos, *An intelligent tutoring system for classifying students into Instructional treatments with mastery scores*

RR-94-13 W.J.J. Veerkamp & M.P.F. Berger, *A simple and fast item selection procedure for adaptive testing*

RR-94-12 R.R. Meijer, *Nonparametric and group-based person-fit statistics: A validity study and an empirical example*

...

Research Reports can be obtained at costs, Faculty of Educational Science and Technology, University of Twente, Mr. J.M.J. Nelissen, P.O. Box 217, 7500 AE Enschede, The Netherlands.



faculty of

EDUCATIONAL SCIENCE AND TECHNOLOGY

A publication by
The Faculty of Educational Science and Technology of the University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

TM030112

NOTICE

REPRODUCTION BASIS

This document is covered by a signed "Reproduction Release (Blanket) form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").