

DOCUMENT RESUME

ED 467 368

TM 034 294

AUTHOR Meijer, Rob R.
TITLE Diagnosing Item Score Patterns Using IRT Based Person-Fit Statistics. Research Report.
INSTITUTION Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.
REPORT NO RR-01-02
PUB DATE 2001-00-00
NOTE 44p.
AVAILABLE FROM Faculty of Educational Science and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
PUB TYPE Reports - Research (143)
EDRS PRICE EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS Adults; *College Students; Diagnostic Tests; Foreign Countries; Goodness of Fit; Higher Education; Intelligence Tests; *Item Response Theory; *Scores; Test Items
IDENTIFIERS *Person Fit Measures; Rasch Model; Test Sensitivity

ABSTRACT

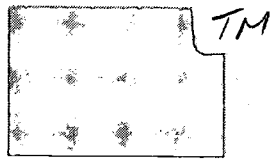
Person-fit statistics have been proposed to investigate the fit of an item score pattern to an item response theory (IRT) model. This study investigated how these statistics can be used to detect different types of misfit. Intelligence test data for 992 people at or beyond college level were analyzed using person-fit statistics in the context of the Rasch model and Mokken's IRT models (R. Mokken, 1997). The sensitivity for different types of misfit was illustrated. The effect of the choice of an IRT model to detect person misfit and the usefulness of person-fit statistics as a diagnostic instrument are discussed. (Contains 1 figure, 7 tables, and 52 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 467 368

Diagnosing Item Score Patterns Using IRT Based Person-Fit Statistics

Research
Report
01-02



Rob R. Meijer

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL HAS
BEEN GRANTED BY

J. Nelissen

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

- This document has been reproduced as received from the person or organization originating it.
- Minor changes have been made to improve reproduction quality.
- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM034294

faculty of
**EDUCATIONAL SCIENCE
 AND TECHNOLOGY**

University of Twente

Department of
Educational Measurement and Data Analysis

BEST COPY AVAILABLE



2

**Diagnosing Item Score Patterns Using
IRT Based Person-Fit Statistics**

Rob R. Meijer

Person-fit statistics have been proposed to investigate the fit of an item score pattern to an item response theory (IRT) model. I investigated how these statistics can be used to detect different types of misfit. Intelligence test data were analyzed using person-fit statistics in the context of the Rasch model and Mokken's IRT models. The sensitivity for different types of misfit was illustrated. The effect of the choice of an IRT model to detect person misfit and the usefulness of person-fit statistics as a diagnostic instrument are discussed.

From the responses to the items on a psychological test a total score is obtained that reflects a person's position on the trait that is being measured. The test score, however, might be inadequate as a measure of a person's trait level. For example, a person may guess some of the correct answers to multiple-choice items on an intelligence test, thus raising his/her total score on the test by luck and not by ability. Similarly, a person not familiar with the test format on a computerized test may obtain a lower score than expected on the basis of his/her ability level. Inaccurate measurement of the trait level may also be caused by sleeping behavior, (e.g., inaccurately answering the first items in a test as a result of, for example, problems of getting started), cheating behavior (e.g., copying the correct answers of another examinee), and plodding behavior (e.g., 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 under the assumption of most test models). Other examples can be found in Wright and Stone (1979, pp. 165-190).

It is important that a researcher has at his or her disposal methods that can help to judge if the item scores of an individual are determined by the construct that is being measured. Person-fit statistics have been proposed that can be used to investigate whether a person answers the items according to the underlying construct the test measures or that other answering mechanisms apply (e.g., Levine & Drasgow, 1985; Meijer & Sijtsma, 1995; 2001; Smith, 1985, 1986; Wright & Stone, 1979). Most statistics are formulated in the context of item response theory (IRT) models (Embretson & Reise, 2000; Hambleton & Swaminathan, 1985) and are sensitive to the fit of an individual score pattern to a particular IRT model. IRT models are useful in describing the psychometric properties of both aptitude and personality measures and are widely used both in psychological and educational assessment.

An overview of the existing person-fit literature (Meijer & Sijtsma, 2001) suggests that most person-fit studies have focused on theoretical development of person-fit statistics and the power of these statistics to detect misfitting item score patterns under varying testing and person characteristics. Most often, simulated data were used that enabled the researcher to distinguish between misfitting and fitting score patterns and thus to determine the power of a person-fit statistic. Also, in most studies a dichotomous decision was made whether the *complete* item score pattern fit or did not fit the IRT model.

However, it may also be useful to obtain information about the subsets of items a person gives unexpected responses, which assumptions of an IRT model have been violated, and how serious the violation is (e.g., Reise, 2000; Reise & Flannery, 1996). Answers to these questions may allow for a more diagnostic approach leading to a better understanding of a person's response behavior. Few studies illustrate systematically the use of these person-fit statistics as a diagnostic tool.

The aim of this paper is to discuss and apply a number of person-fit statistics proposed for parametric IRT and nonparametric IRT models. In particular, I will apply person-fit statistics in the context of the Rasch (1960) model and Mokken's (1971; 1997) IRT models that can be used to diagnose different kinds of aberrant response behavior. Within the context of both kinds of IRT models, person-fit statistics have been proposed that can help the researcher to diagnose the item scores. Existing studies, however, are conducted using either a parametric or nonparametric IRT model, and it is unclear how parametric and nonparametric person-fit statistics relate to each other. To illustrate the diagnostic use of person-fit statistics I will use empirical data from an intelligence test in the context of personnel selection.

This paper is organized as follows. First, I will introduce the basic principles of IRT and discuss parametric and nonparametric IRT. Because nonparametric IRT models are relatively unknown, I will discuss nonparametric models more extensively. Second, I will introduce person-fit statistics and person tests that are sensitive to different types of misfitting score patterns for both parametric and nonparametric IRT modeling. Third, I will illustrate the use of the statistics to study misfitting item score patterns associated with misunderstanding instruction, item disclosure, and random response behavior using the empirical data of an intelligence test. Finally, I will give the researcher some suggestions as to which statistics can be used to detect specific types of misfit.

IRT models

Parametric IRT

Fundamental to IRT is the idea that psychological constructs are latent, that is, not directly observable and that knowledge about these constructs can only be obtained through the

manifest responses of persons to a set of items. IRT explains the structure in the manifest responses by assuming the existence of a latent trait (θ) on which persons and items have a position. IRT models allow the researcher to check if the data fit the model. The focus in this article is on IRT models for dichotomous items. Thus, one response category is positively keyed (item score 1), whereas the other is negatively keyed (item score 0); for ability and achievement items these response categories are usually "correct" and "incorrect", respectively.

In IRT, the probability of obtaining a correct answer on item g ($g = 1, \dots, k$) is a function of θ and characteristics of the item. This conditional probability $P_g(\theta)$ is the item response function (IRF). Item characteristics that are often taken into account are the item discrimination (a), the item location (b), and the pseudo-chance level parameter (c). The item location b is the point at the trait scale where the probability of a correct response is 0.5. The greater the value of the b parameter, the greater the ability that is required for an examinee to have a 50% chance of correctly answering the item; thus the harder the item. Difficult items are located to the right or the higher end of the ability scale; easy items are located to the left of the ability scale. When the ability levels are transformed so their mean is 0 and their standard deviation is 1, the values of b vary typically from about -2 (very easy) to +2 (very difficult). The a parameter is proportional to the slope of the IRF at the point b on the ability scale. In practice, a ranges from 0 (flat IRF) to 2 (very steep IRF). Items with steeper slopes are more useful for separating examinees near an ability level θ . The pseudo-chance level parameter c (ranging from 0 to 1) is the probability of a 1 score for low-ability examinees (that is, $\theta \rightarrow -\infty$).

In parametric IRT, $P_g(\theta)$ often is specified using the 1-, 2-, or 3-parameter logistic model (1-, 2-, 3PLM). The 3PLM (Lord & Novick, 1968, Chaps. 17-20) is defined as

$$P_g(\theta) = c_g + \frac{(1 - c_g) \exp[a_g(\theta - b_g)]}{1 + \exp[a_g(\theta - b_g)]}. \quad (1)$$

The 2PLM can be obtained by setting $c_g = 0$ for all items; and the 1PLM or Rasch (1960) model can be obtained by additionally setting $a_g = 1$ for all items. In the 2- and 3PLM the IRFs may cross, whereas in the Rasch model the IRFs do not cross. An advantage of the Rasch model is that the item ordering according to the item difficulty is the same for each θ value, which facilitates the interpretation of misfitting score patterns across θ .

Most IRT models assume unidimensionality, and a specified form for the IRF which can be checked empirically. Unidimensionality means that the latent space that explains the person's test performance is unidimensional. Related to unidimensionality is the assumption of local independence. Local independence states that the responses in a test are statistically independent conditional on θ . Thus, local independence is evidence for unidimensionality if the IRT model contains person parameters on only one dimension.

For this study it is important to understand that IRT models are stochastic versions of the deterministic Guttman (1950) model. The Guttman model is defined by

$$\theta < b_g \leftrightarrow P_g(\theta) = 0 \quad (2)$$

and

$$\theta \geq b_g \leftrightarrow P_g(\theta) = 1. \quad (3)$$

The model thus excludes a correct answer on a relatively difficult item and an incorrect answer on an easier item. The items answered correctly are always the easiest or most popular items on the test. These principles are not restricted to items concerning knowledge, but also apply to the domains of intelligence, attitude, and personality measurement.

This view of test behavior leads to a deterministic test model, in which a person should never answer negatively to an item when he/she answers correctly to a more difficult item. An important consequence is that given the total score, the individual item responses can be reproduced. On the person level this implies the following. When I assume throughout this paper that the items are ordered from easy to difficult, it is expected on the basis of the Guttman model that given a total score X_+ , the correct responses are given on the first X_+ items and the incorrect responses are given on the remaining $k - X_+$ items. Such a pattern is called a Guttman pattern. A pattern with all correct responses in the last X_+ positions and incorrect responses in the remaining positions is called a reversed Guttman pattern. As Guttman (1950, p. 64) observed, empirically obtained test data are often not perfectly reproducible. In IRT models, the probability of answering an item correctly is

between 0 and 1 and thus errors in the sense that an easy item is answered incorrectly and a difficult item is answered correctly are allowed. Many reversals, however, point to aberrant response behavior.

Parametric Person-Fit Methods

Model-data fit can be investigated on the item or person level. Examples of model-data fit studies on the item level can be found in Thissen and Steinberg (1988), Meijer, Sijtsma, and Smid (1990), and Reise and Waller (1993). Because the central topic in this study is person fit, I will not discuss these item-fit statistics. The interested reader should refer to the references above for more details.

Although several person-fit statistics have been proposed, I will discuss two types of fit statistics that can be used in a complementary way: (a) statistics that are sensitive to violations against the Guttman model and (b) statistics that can be used to detect violations against unidimensionality. I will use statistics that can be applied without modeling an alternative type of misfitting behavior (Levine & Drasgow, 1988). Testing against a specified alternative is an option when the researcher knows what kind of misfit to expect. This approach has the advantage that the power is often higher than when no alternative is specified. The researcher, however, is often unsure about the kind of misfit to expect. In this situation, the statistics discussed below are useful.

There are several person-fit statistics that can be applied using the Rasch model. I illustrate two methods that have well-known statistical properties. One of the statistics is a uniformly most powerful test (e.g., Lindgren, 1993, p. 350).

Violations against the Guttman model. Most person-fit methods in the context of the 2- and 3PLM are based on determining the likelihood of a score pattern. Many studies have been conducted using the log-likelihood function l . Let X_g denote the item score on item g , then

$$l = \sum_{g=1}^k \{X_g \ln P_g(\theta) + (1 - X_g) \ln[1 - P_g(\theta)]\}. \quad (4)$$

A standardized version of this statistic is denoted l_z (Drasgow, Levine, and Williams, 1985; Levine and Rubin, 1979). l_z was proposed to obtain a statistic that was less confounded with θ than l , that is, a statistic whose value is less dependent on θ . l_z is

given by

$$l_z = \frac{l - E(l)}{\sqrt{Var(l)}}, \quad (5)$$

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

$$E(l) = \sum_{g=1}^k \{P_g(\theta) \ln [P_g(\theta)] + [1 - P_g(\theta)] \ln [1 - P_g(\theta)]\}, \quad (6)$$

and

$$Var(l) = \sum_{g=1}^k P_g(\theta) [1 - P_g(\theta)] \left[\ln \frac{P_g(\theta)}{1 - P_g(\theta)} \right]^2. \quad (7)$$

Large negative l_z values indicate aberrant response behavior. Below I will give an example of the use of l in the context of the Rasch model.

To classify an item score pattern as (mis)fitting, a distribution is needed under response behavior that fit the IRT model. For long tests (larger than, say, 80 items) and true θ it can be shown that l_z is distributed as standard normal. A researcher can specify a type I error rate, say $\alpha = 0.05$, and classify an item score pattern as misfitting when $l_z < -1.65$. In practice, however, θ must be estimated and with short tests this leads to thicker tail probabilities than expected under the standard normal distribution which results in a conservative classification of item score patterns as misfitting (Nering, 1995, 1997; Molenaar & Hoijtink, 1990). Therefore, Snijders (2001) derived an asymptotic sampling distribution for a family of person-fit statistics like l where a correction factor was used for the estimate of θ , denoted $\hat{\theta}$. For an empirical example using this correction factor see Meijer and van Krimpen-Stoop (2001).

The l or l_z statistic is most often used in the context of the 2PLM or 3PLM. Because I use the Rasch model to analyze an empirical dataset, a simplified version of l is employed. For the Rasch model, l can be simplified as the sum of two terms,

$$l = d + M \quad (8)$$

with

$$d = - \sum_{g=1}^k \ln[1 + \exp(\theta - b_g)] + \theta X_+ \quad (9)$$

and

$$M = - \sum_{g=1}^k b_g X_g. \quad (10)$$

Given $X_+ = \sum_{g=1}^k X_g$ (that is, given $\hat{\theta}$, which in the Rasch model only depends on the sufficient statistic X_+), d is independent of the item score pattern (X_g is absent in Equation 9) and M is dependent on it. Given X_+ , l and M have the same ordering in the item score pattern, that is, the item score patterns are ordered similarly by M and l . Because of its simplicity, Molenaar and Hoijtink (1990) used M rather than l as a person-fit statistic.

To illustrate the use of M consider all possible item score patterns with $X_+ = 2$ on a five-item test. In Table 1 all possible item score patterns on the test are given with their M -values assuming $b = (-2, -1, 0, 1, 2)$. Pattern #1 is most plausible, that is, this pattern has the highest probability under the Rasch model. Pattern #10 is least plausible under this configuration of item difficulties. Note that the items difficulty parameters as explained above are centered around zero.

Molenaar and Hoijtink (1990) proposed three approximations to the distribution of M : (1) complete enumeration, (2) a chi-square distribution, where the mean, standard deviation, and skewness of M are taken into account, and (3) a distribution obtained via Monte Carlo simulation. For all scores complete enumeration was recommended for tests with $k \leq 8$, and up to $k = 20$ in the case of the relatively extreme scores $X_+ = 1, 2, k - 2, k - 1$. For other cases a chi-square distribution was proposed with the exception of very long tests for which Monte Carlo simulation was recommended. For a relatively simple introduction to this statistic and a small application see Molenaar and Hoijtink (1995).

Insert Table 1 about here

Statistics like l_z and M are often presented as statistics to investigate the general fit of an item score pattern. However, note that l and M are sensitive to a specific type of misfit

to the IRT model; namely, violations to Guttman patterns. As I discussed above, when the items are ordered from easy to difficult, an item score pattern with correct responses in the first X_+ positions and incorrect responses in the remaining $k - X_+$ positions is called a Guttman pattern because it meets the requirements of the Guttman (1950) model. To illustrate this, consider again the item score patterns in Table 1. The pattern of person #1 is a perfect Guttman pattern which results in the maximum value of M and the pattern of person #10 is the reversed Guttman pattern which results in a minimum value of M .

As an alternative to M , statistics discussed by Wright and Stone (1979) and Smith (1986) can be used. For example, the statistic

$$B = \sum_{g=1}^k \frac{(X_g - P_g(\theta))^2}{kP_g(\theta)(1 - P_g(\theta))}. \quad (11)$$

I prefer the use of M because research has shown (e.g., Molenaar & Hoijtink, 1990) that the actual type I error rates are too sensitive for this statistic to the choice of the θ distribution, item parameter values, and θ level to be trusted, and the advocated standardizations are, in most cases, unable to repair these deficiencies. For example, Li and Olejnik (1997) concluded that for both l_z and a standardized version of B the sampling distribution under the Rasch model deviated significantly from the standard normal distribution. Also Molenaar and Hoijtink (1995) found for a standardized version of B and a standard normal distribution for θ using 10,000 examinees that the mean of B was -0.13, the standard deviation was 0.91 and the 95% percentile was 1.33 rather than 1.64 and that thus too few respondents would be flagged as possible aberrant.

Violations against unidimensionality. To investigate unidimensionality I will check whether the ability parameters are invariant over subtests of the total test. In many cases, response tendencies that lead to deviant item score patterns cause violations of unidimensional measurement, that is, violations against the invariance hypothesis of θ for a suitably chosen subdivision of the test. Trabin and Weiss (1983) discuss, in detail, how factors like carelessness, guessing, or cheating may cause specific violations against unidimensional measurement when the test is subdivided into subtests of different difficulty level. It is also useful to investigate other ways of subdividing the test. For example, the order of presentation, the item format, and the item content

provide potentially interesting ways of subdivision. Each type of split can be used to extract information concerning different response tendencies underlying deviant response vectors.

Using the Rasch model, Klauer (1991) proposed a uniformly most powerful test that is sensitive to the violation of unidimensional measurement. The statistical test is based on subdivisions of the test into two subtests, A_1 and A_2 . Let the scores on these two subtests be denoted by X_1 and X_2 and let the latent traits that underlie the person's responses on these two subtests be denoted by θ_1 and θ_2 , respectively. Under the null hypothesis (H_0), it is assumed that the probabilities underlying the person's responses are in accordance with the Rasch model for each subtest. Under the alternative hypothesis (H_1), the two subtests need not measure the same latent trait, and different ability level θ_k may underlie the person's performance with respect to each subtest A_k . An individual's deviation from invariance is given by the parameter $\eta = \theta_1 - \theta_2$ and $H_0 : \eta = 0$ is tested against $H_1 : \eta \neq 0$. To test this hypothesis Klauer (1991) considered the joint distribution of X_1 and a person's total score, X_+ . This joint distribution is determined on the basis of the Rasch model. Given a prespecified nominal type I error, cut-off scores are determined for subtest scores X_1 .

Klauer (1991) gives an example for a test consisting of 15 items with difficulty parameters between -0.69 and 0.14, where the test is divided into two subtests of seven easy and eight difficult items. Let the cut-off scores be denoted by $c_L(X_+)$ and $c_U(X_+)$. For a test score of, for example, $X_+ = 8$ the cut-off scores for X_1 are $c_L(X_+) = 3$ and $c_U(X_+) = 6$. Thus, score patterns with $X_1 = 3, 4, 5, \text{ or } 6$ are considered to be consistent with the Rasch model. A value of X_1 outside this range points at deviant behavior. Thus if an examinee with $X_+ = 8$ has only 2 correct answers on the first subtest ($X_1 = 2$) and 6 correct answers on the second subtest ($X_2 = 6$), the item score pattern will be classified as aberrant. Also, note that an item score pattern with $X_1 = 7$ and $X_2 = 1$ will be classified as aberrant. For this score pattern there are too many correct answers on the first subtest and too few correct answers on the second subtest given the stochastic nature of the Rasch model.

The test statistic equals

$$X^2 = -2 \log(p(X_k|X_+)) \quad (12)$$

where X_k is the subtest score and $p(X_k|X_+)$ is the conditional probability discussed above of observing deviations this large or even larger from the invariance hypothesis. Under the Rasch model X^2 is chi-squared distributed with two degrees of freedom.

To illustrate the use of this statistic I consider an empirical example given by Klauer (1991). A verbal analogies test consisting of 47 items was divided into two subtests of 23 easy and 24 difficult items. Then a person with a total score of 10 but only 3 correct answers on the easy subtest obtained $X^2 = 17.134$ with $p(X_k|X_+) = 0.0003$. For this person, the invariant hypothesis was violated. Power curves for statistics analogous to X^2 can be found in Klauer (1995).

Note that the statistic X^2 and M differ on the types of violations to the Rasch model. The observed item score patterns is considered inconsistent with the Rasch model if $M \leq c(X_+)$, where $c(X_+)$ is a cut-off score that depends on the test score X_+ associated with the response vector and on the chosen α -level. For example, assume that in Table 1 $c(X_+) = -2$, then item patterns #9 and #10 are classified as inconsistent with the Rasch model. Assume now as an alternative hypothesis, H_1 , that a few examinees exhibit misfitting test behavior described above as "plodding" behavior, which probably result in an almost perfect Guttman pattern like pattern #1. In this case maximal values of the statistic M are obtained, indicating perfect fit. In contrast, the statistic X^2 will flag such a pattern as misfitting when I split the test into a subtest with the first $k/2$ items and a subtest with the second $k/2$ items (assuming an even number of items). This pattern is inconsistent with the Rasch model because X_1 on the easiest subtest is too high and the X_2 on the second subtest is too low given the assumptions of the (stochastic) Rasch model.

Important is that in this study I will calculate X^2 on the basis of the item difficulty ordering in the test, X^2 will then be denoted by X_{dif}^2 , and on the basis of the presentation order of the items in the test, denoted X_{ord}^2 .

Nonparametric IRT

Although parametric models are used in many IRT applications, nonparametric IRT

models are becoming more popular (e.g., Stout, 1990). For a review of nonparametric IRT, see Sijtsma (1998) or Junker and Sijtsma (2001). In this study I will analyze the data by means of the Mokken (1971) models. I use these models because they are popular nonparametric IRT models (e.g., Mokken, 1997; Sijtsma, 1998). There is also a user friendly computer program, MSP5 for windows, to operationalize these models (Molenaar & Sijtsma, 2000).

The first model proposed by Mokken (1971; 1997; see also Molenaar, 1997) is the model of monotone homogeneity (MHM). This model assumes unidimensional measurement and an increasing IRF as function of θ . However, unlike parametric IRT, the IRF is not parametrically defined and the IRFs may cross. The MHM allows the ordering of persons with respect to θ using the unweighted sum of item scores. In many testing applications, it often suffices to know the order of persons on an attribute (e.g., in selection problems). Therefore, the MHM is an attractive model for two reasons. First, ordinal measurement of persons is guaranteed when the model applies to the data. Second, the model is not as restrictive with respect to empirical data as the Rasch model and thus can be used in situations where the Rasch model does not fit the data.

The second model proposed by Mokken (1971) is the model of double monotonicity (DMM). The DMM is based on the same assumptions as the MHM plus the additional assumption of nonintersecting IRFs. Under the DMM it is assumed that when k items are ordered and numbered, the conditional probabilities of obtaining a positive response are given as,

$$P_1(\theta) \geq P_2(\theta) \geq \dots \geq P_k(\theta), \text{ for all } \theta. \quad (13)$$

Thus the DMM specifies that, except for possible ties, the ordering is identical for all θ values. Note that ordering of both persons and items is possible when the DMM applies, however, attributes and difficulties are measured on separate scales. Attributes are measured on the true score scale, and difficulties are measured on the scale of proportions. Thus, persons can be ordered according to their true scores using the total score. For the Rasch model, measurement of items and persons takes place on a common difference or a common ratio scale.

An important difference between the Mokken and the Rasch models is that the IRFs for the Mokken models need not be of the logistic form. This difference makes Mokken models less restrictive to empirical data than the Rasch model. Thus the Mokken models can be used to describe data that do not fit the Rasch model.

In Figure 1 examples are given of IRFs that can be described by the MHM, DMM, and the Rasch model.

Insert Figure 1 about here

This difference also suggests that an item score pattern will be less easily classified as misfitting under the MHM than under the Rasch model because, under a less restrictive model, item score patterns are allowed that are not allowed under a more restrictive model. To use an extreme example, assume the data are described by the deterministic Guttman model. In the case of the Guttman model every item score pattern with an incorrect score on an easier item and a correct score on a more difficult item will be classified as misfitting.

Nonparametric person-fit statistics

Person-fit research using nonparametric IRT has been less popular than with the parametric IRT modeling. Although several person-fit statistics have been proposed which may be used in nonparametric IRT modeling, Meijer and Sijtsma (2001) conclude that one of the drawbacks of these statistics is that the distribution of the values of these statistics is dependent on the total score and that the distributional characteristics of these statistics conditional on the total score is unknown.

Van der Flier (1982) proposed a nonparametric person-fit statistic U_3 that was purported to be normally distributed (see e.g., Meijer, 1998). Recently, Emons, Meijer, and Sijtsma (in press) investigated the characteristics of this statistic and concluded that in most cases studied there was a large discrepancy between the type I error rate under the theoretical sampling distribution and the empirical sampling distribution. This finding makes it difficult to interpret person-fit scores based on this statistic.

Another nonparametric person-fit method was proposed by Sijtsma and Meijer (2001). They discussed a statistical method based on the person response function (PRF, Trabin & Weiss, 1983; Reise, 2000). The PRF describes the probability that a respondent

with fixed θ correctly answers items from a pool of items with varying location. For the Rasch model the PRF can be related to the IRF by changing the roles of θ and b where b is treated as a random item variable and θ is treated as fixed. The PRF is a nonincreasing function of item difficulty, that is, the more difficult the item the smaller the probability that a person answers an item correctly. Under the DMM, local deviations from nonincreasingness can be used to identify misfit.

To detect local deviations an item score pattern is divided into G subtests of items, so that A_1 contains the m easiest items, A_2 contains the next m easiest items, and so on. Let subtests A_g of m increasingly more difficult items be collected in mutually exclusive vectors A_g such that $A = (A_1, A_2, \dots, A_G)$. Consider newly defined vectors $A_{(g)}$ each of which contains two adjacent subsets A_g and A_{g+1} : $A_{(1)} = (A_1, A_2)$, $A_{(2)} = (A_2, A_3)$, ..., $A_{(G-1)} = (A_{G-1}, A_G)$. The statistical method applies to each pair in $A_{(g)}$. A useful question in person-fit research is whether the number of correct answers on the more difficult items (denoted X_{+d}) is exceptionally low given the total score and the subtest score on the easier subtest (denoted X_{+e}). To test this, Sijtsma and Meijer (2001; see also Rosenbaum, 1987) showed that for each pair of subtest scores a conservative bound based on the hypergeometric distribution can be calculated (denoted \mathcal{P}) for the probability that a person has, at most, $X_{+e}1s$ on the easiest subtest. If, for a particular pair, this probability is lower than, say 0.05, then the conclusion is that X_{+e} in this pair is unlikely given X_{+d} and X_+ . A conservative bound means that the probability under the hypergeometric distribution is always equal or larger than under a nonparametric IRT model. Thus, if $\mathcal{P} = 0.06$ is found, then under the IRT model this probability is smaller than 0.06. Although this method is based on item response functions that are nonintersecting, Sijtsma and Meijer (2001) showed that the results are robust against violations of this assumption. Furthermore, they investigated the power of \mathcal{P} to detect careless response behavior. Detection rates ranged from .018 through .798 depending on the θ level.

To illustrate the use of this method assume that the items are ordered from easy to difficult and consider a person who generates the following item score pattern (11110|00000). This person has four correct answers on the five easiest items which is as expected under the DM model and the cumulative probability under the hypergeometric

distribution equals 1.00. However, if I consider a pattern (00000|10111) then this cumulative probability equals .0238 and it is concluded that at a 5% level this item score pattern is unlikely. Note that this method takes into account the subtest score, and not the different permutations of zeros and ones within a subtest. In the Appendix the calculations are given.

Sijtsma and Meijer (2001) did not relate \mathcal{P} to specific types of misfit. However, if I compare this method with M and X^2 , it can be concluded that \mathcal{P} is sensitive to violations of unidimensionality when the test is split into subtests according to the item difficulty and for subtests with large values of X_{+d} in combination with small values of X_{+e} . However, it is insensitive to violations of unidimensionality when X_{+e} is larger than X_{+d} . Also, random response behavior will not be detected easily because the expected X_+ is approximately the same on each subtest, which will not result in small values of \mathcal{P} . Thus, the method proposed by Sijtsma and Meijer (2001) can be considered a special case within a nonparametric IRT framework of the method proposed by Klauer (1991), as previously discussed.

Method

To illustrate the usefulness of the statistics to detect different types of misfit I used empirical data of the Test for Nonverbal Abstract Reasoning (TNVA, Drenth, 1969). The test consist of 40 items and the test is speeded. A sample of 992 persons was available. The test is developed for persons at or beyond college/university level. For the present analysis the sample consisted of high school graduates. The dataset was collected for personnel selection in computer occupations, such as systems analyst and programmer. I first investigated the fit of this dataset using the computer program RSP (Glas and Ellis, 1993) and MSP5 for Windows (Molenaar & Sijtsma, 2000). The program RSP contains fit statistics for monotonicity and local independence that are improved versions of the statistic proposed by Wright and Panchepakesan (1969; see also Smith, 1994). It also contains statistics for testing the null hypothesis that the IRFs are logistic with equal slopes against the alternative that they are not. I concluded that the items reasonably fit the Rasch model and the Mokken models. Details about this fit procedure can be obtained by the author. Before I analyze the item score patterns, I will first discuss some possible types

of person misfit that can occur on the TNVA.

Types of Possible Misfit in Intelligence Testing

Misunderstanding of instruction/Using a suboptimal strategy.

To maximize performance on the TNVA, examinees must work quickly because there is a limited amount of time to complete the test. However, some examinees may focus on accuracy, afraid of answering an item incorrectly and, as a result, may generate a pattern where the answers on the first items are (almost) all correct and on the later items are (almost) all incorrect ("plodding" behavior as discussed in Wright & Stone, 1979). This strategy is especially problematic for the TNVA because the first items are relatively difficult (for example item #5 has a proportion-correct score equal to 0.55). A person working very precisely will spend too much time answering the first items and the total score on the test may then underestimate his/her true score.

As a result of this type of misfit it is expected that given X_+ the test score on the first part of the test is too large and on the second part too low given the stochastic nature of an IRT model. X_{ord}^2 can be used to detect this type of misfit whereas M is less suitable. X_{dif}^2 will only be sensitive to this type of misfit when the item presentation ordering matches the item difficulty ordering in the test. \mathcal{P} is not useful here. Remember that \mathcal{P} assumes that the items are ordered from easy to difficult and that the method is sensitive to unexpected low subtest scores on an easy subtest in combination with high subtest scores on a more difficult subtest. Working very precisely, however, will result in the opposite pattern; namely, high subtest scores on the easy subtest and low subtest scores on the more difficult subtest.

Item Disclosure

When a test is used in a selection situation with important consequences for individuals, persons may be tempted to obtain information about the type of test questions or even about the correct answers to particular items. In computerized adaptive testing this is one of the major threats to the validity of the test scores, especially when items are used repeatedly. This is also a realistic problem for paper-and pencil tests. The TNVA is often used in the Netherlands by different companies and test agencies because there are no

alternatives. Item exposure is therefore realistic and may result in a larger percentage of correct answers than expected on the basis of the trait that is being measured.

Detection of item exposure by means of a person-fit statistic is difficult because for a particular person it is often unknown which items are known and how many items are known. Advance knowledge of a few items will only have a small effect on the test score and is therefore not so easily detected. Also, advance knowledge of the easiest items will only have a small effect on the test score. This outcome suggests that advance knowledge of the items of median difficulty and of the most difficult items may have a substantial effect on the total score, especially for persons with a low θ level. This type of misfit may be detected by fit statistics like M , X_{dif}^2 , and \mathcal{P} . Note that X_{ord}^2 will only be sensitive to item disclosure when the item difficulty order matches the order of presentation in the test.

Note that item disclosure has some similarities with response distortion on personality scales (Reise & Waller, 1993). Although the mechanisms beyond those two types of misfit are different, the realized item scores patterns may be similar: both are the result of faking and both have the effect of increasing the test score. Recently, Zickar and Robie (1999) investigated the effect of response distortion on personality scales for applicants in selection situations. Response distortion occurs when people present themselves in a favorable light so that they make a more favorable impression, for example, in selection situations where personality tests are used to predict job success, job effectiveness, or management potential. Zickar and Robie (1999) examined the effects of experimentally-induced faking on the item- and the scale-level measurement properties of three personality constructs that possess potential utility for use in personnel selection (Work orientation, Emotional Stability, and Nondelinquency). Using a sample of military recruits, comparisons were made between an honest condition and both an ad-lib (i.e., "fake good" with no specific instructions on how to fake) and coached faking (i.e., "fake good" with instructions on how to "beat" the test) condition. Faking had a substantial effect on the personality latent trait scores and resulted in a moderate degree of differential item functioning and differential test functioning. These results suggests that identifying faking is important.

Random response behavior

Random response behavior may be the result of different underlying mechanisms depending on the type of test and the situation in which the test is administered. One reason for random response behavior is lack of motivation. Another reason is lack of concentration which may result from feeling insecure. In intelligence testing random response behavior may occur as a result of answering questions too quickly. X_{dif}^2 , X_{ord}^2 and M may be sensitive to random response behavior, depending on which items are answered randomly. \mathcal{P} seems to be less suited to detect random response behavior because \mathcal{P} is sensitive to unexpected low subtest scores on an easy subtest in combination with high subtest scores on a more difficult subtest. Random response behavior as a result of answering questions too quickly will probably result in similar number-correct scores for different subtests.

In Table 2 the sensitivity of the statistics to different types of deviant response behavior is depicted, where a "+" sign indicates that the statistic is sensitive to the response behavior and a "-" sign indicates that the statistic is expected to be insensitive to the response behavior. The reader, however, should be careful when interpreting the "+" and "-" signs. For a particular type of deviant behavior, the sensitivity of a statistic depends on various factors, like, for example, whether the presentation ordering of the items is in agreement with the difficulty order. If this is the case both X_{dif}^2 and X_{ord}^2 are sensitive to "Misunderstanding of Instruction".

Insert Table 2 about here

To illustrate the validity of the different fit statistics I conducted a small simulation study. 10,000 item score patterns were simulated on a 40-items test using the Rasch model with item difficulties drawn from a uniform distribution between [-2,2] and θ drawn from a standard normal distribution. The three types of aberrant response behavior were simulated as follows. The use of a suboptimal strategy was simulated by changing the item scores on the first 10 items: the probability of responding correctly to these items was .90, whereas the probability equalled .10 for the remaining items. Item disclosure was simulated by changing the item scores on the 5 most difficult items, the probability of answering an item correctly was .90, and random response behavior was simulated by

changing the answers on the 10 most difficult items so that the probability of answering an item correctly equalled .20. The proportion correctly identified misfitting score patterns for the different statistics is given in Table 3. As expected X_{ord}^2 had

Insert Table 3 about here

the highest power for detecting a suboptimal strategy, whereas \mathcal{P} was insensitive to this type of response behavior. X_{ord}^2 was insensitive to item disclosure and \mathcal{P} was insensitive to random response behavior.

Procedure

Person-fit statistics

The program RSP (Glas & Ellis, 1993) was used to compute the p -values (significance probabilities) belonging to the M -statistic (Equation 10). A program developed by Klauer (1991) was used to calculate the X^2 person-fit statistic and a program developed in S-plus (2000) was used to determine \mathcal{P} . This program can be obtained from the author.

In the person-fit literature there has not been much debate about the nominal type I error rate. In general, I prefer to choose a relatively large α , for example $\alpha = .05$ or $\alpha = .10$, for two reasons. The first reason is that existing person-fit tests have relatively low power due to limited test length and a relatively small number of observations. Choosing a small value for α (e.g., $\alpha = .001$) will then result in a very low number of misfitting score patterns. The second reason is that, in practice, persons will never be withheld a total score solely on the basis of the fit of their score pattern to an IRT model (i.e., having an extreme person-fit score). A person-fit statistic will alert the researcher that a person's response behavior is unexpected and that it may be interesting to study the item score pattern more closely. This implies that, in practice, incorrectly rejecting the null hypothesis of fitting response behavior has no serious consequences. In this study I will use $\alpha = 0.10$, $\alpha = 0.05$, and $\alpha = 0.01$.

Results

Descriptive statistics

The mean number-correct score on the TNVA was 27.08 with a standard deviation of 5.73. There were no persons who answered all items incorrectly and there was one person who answered all 40 items correctly. The item proportion correct scores varied between 0.11 (item 40) through 0.99 (item 4).

Person-Fit results

Items ordered according to the item difficulty.

Table 4 shows the proportion of score patterns classified as misfitting (detection rates) for M , X_{dif}^2 , X_{ord}^2 , and \mathcal{P} . The results using X_{ord}^2 will be discussed in the next paragraph. Note that M is based on the complete item score

Insert Table 4 about here

pattern, whereas the X^2 fit statistics and \mathcal{P} are based on a split of the item score patterns into subtests. By means of X^2 and \mathcal{P} , the number-correct scores on two subtests are compared. I first split the test into a subtest containing the 20 easiest items and a subtest containing the 20 most difficult items. Table 4 shows that for M the proportion of misfitting item score patterns was somewhat higher than the nominal α -levels in all conditions. For the X_{dif}^2 statistic the proportions at $\alpha = 0.05$ and $\alpha = 0.10$ were higher than for the M -statistic.

To calculate \mathcal{P} I first ordered the items according to decreasing item proportion correct score and split the test into two subtests of 20 items each. Table 4 shows that no item score pattern was classified as misfitting. Therefore, I split the test into 4 subtests of 10 items each, where subtest #1 contained the easiest items, subtest #2 contained the next 10 easiest items, and so on. This was done because Sijtsma and Meijer (2001) showed empirically that the power of \mathcal{P} often increases when smaller subtests are used. Calculating \mathcal{P} using subtest #1 and subtest #2, however, no significant results were obtained. Thus, the

proportion of item score patterns classified as misfitting was much higher using X_{dif}^2 than using \mathcal{P} . For example, for $\alpha = 0.05$ using X_{dif}^2 , I found a proportion of 0.093 and for $\alpha = 0.10$ I found a proportion of 0.149. \mathcal{P} is only sensitive to score patterns with unexpected 0 scores on the easier subtest in combination with unexpected 1 scores on the more difficult subtest. Therefore, I also determined the detection rates using X_{dif}^2 when the total score on the second subtest was unexpectedly high compared to the second subtest. For the split into two subtest of 20 items and $\alpha = 0.01$, I found a proportion of 0.012; for $\alpha = 0.05$ a proportion of 0.043 and for $\alpha = 0.10$ a proportion of 0.051. Although the proportions are decreasing for this particular type of misfit, X_{dif}^2 still classifies a higher proportion of item score patterns as misfitting than \mathcal{P} .

In Table 5 the four most deviant score patterns according to the X_{dif}^2 test are given with their corresponding p -values for the X_{dif}^2 test and the M -statistic. Person # 754 with $X_+ = 12$ produces a very unexpected item

Insert Table 5 about here

score pattern because only 4 out of the 20 easiest items are answered correctly, whereas on the 20 most difficult items 8 items are answered correctly. This pattern was also classified as unexpected using M . To further illustrate this unexpected behavior, item #1 with a proportion correct score of 0.99 was answered incorrectly, whereas item #39 with a proportion correct score of 0.19 was answered correctly. Consider the types of aberrant response behavior discussed above to interpret the type of misfit. It is clear that "plodding" and "item disclosure" are very unlikely and that this response pattern may be the result of "random response behavior". Note that the expected X_+ for a person who is completely guessing on this test with 40 items with four alternatives per item equals 10, almost equal to the total score of person #754 with $X_+ = 12$.

The pattern of person #754 was classified as *fitting* using \mathcal{P} . Dividing the test into two subtests of 20 items each resulted in $\mathcal{P} = 0.150$. Also dividing the test into four subtests of 10 items each and comparing the first two subtests resulted in a nonsignificant result: $\mathcal{P} = 0.500$.

The difference between the X^2 statistics and M can be illustrated by the item score pattern of person # 841. This person with $X_+ = 19$, answers 19 out of the 20 easiest

items correctly, and answers the 20 most difficult items incorrectly. This is unexpected on the basis of the probabilistic nature of the Rasch model. Thus I expect less correct scores on the easier subtest and more correct answers on the second subtest given $X_+ = 19$. For this pattern the X_{dif}^2 test has $p = 0.0004$. However, for the M statistic $p = 0.4521$ and $\mathcal{P} = 1.000$. This illustrates that both M and \mathcal{P} are *insensitive* and X_{dif}^2 is *sensitive* to violations of unidimensionality. To interpret this response behavior consider also the value of X_{ord}^2 because this type of answering behavior has many similarities with "plodding behavior". For person # 841 using X_{ord}^2 $p = 0.0019$. Thus on the basis of the item ordering, this pattern is very unlikely. This pattern is a good example of someone who is working very slowly but precisely, and uses a suboptimal strategy to obtain a maximal score given his/her ability. If he/she had guessed the answers on the more difficult items later in the test, the test score would have been higher.

Note that the item score pattern of person #669 is also very unlikely. This person has 14 correct answers on the first 20 items and 14 correct answers on the second most difficult items. Thus, given $X_+ = 27$, too many correct answers on the difficult items and too few correct answers on the easier items. This pattern is more difficult to interpret than the other examples when only relying on the item difficulty order. Using X_{ord}^2 $p = 0.0141$; thus on the basis of the item presentation order this pattern is less unlikely. Item preview of the more difficult items may be an explanation, or it may be someone who is working very fast and easily skips items.

Items ordered according to the rank order of presentation in the test

Different types of score patterns may be classified as misfitting depending on whether I use the presentation order on the test or the item difficulty order. I will only discuss the difference between X_{dif}^2 and X_{ord}^2 because M is not influenced by the ordering of the items, and \mathcal{P} is only based on the item difficulty order.

Because the order of the items in the TNVA according to presentation and difficulty order are partly overlapping, some item score patterns are classified as misfitting by both X^2 statistics (see below). However, there are exceptions. In Table 6, it can be seen that person #5, #827, and #101 produce item score patterns with many

Insert Table 6 about here

correct item scores on the first subtest and many incorrect scores on the second subtest. The ratio of the test scores on the two subtests for these persons was 16/0, 20/6, and 19/2. This indicates people who work slowly but conscientiously. Note, that for these people, $X_{dif}^2 p < .01$, whereas using $X_{ord}^2 p > 0.05$. When I combine the information obtained from both statistics I conclude that these people answer both difficult items and easy items correctly (if not, they would have had very low p -values using X_{dif}^2) and mainly answer the first items in the test correctly. This is strong evidence for "plodding" behavior.

Overlap of the statistics

In Table 7 the overlap of the number of item score patterns classified as misfitting at $\alpha = 0.05$ by the different person-fit statistics is given. This table illustrates that the three

Insert table 7 about here

statistics are sensitive to different types of misfit. Note, however, that of the 69 patterns classified as misfitting by M only 22 are also classified as misfitting by X_{ord}^2 and only 19 are classified as misfitting using X_{dif}^2 . For X^2 , using the item difficulty- or presentation order has an effect: the overlap of the number of items score patterns classified as misfitting is only 30, whereas 88 patterns are classified as misfitting using X_{ord}^2 and 93 patterns are classified as misfitting using X_{dif}^2 .

Discussion

In this study, the item score patterns on an intelligence test were analyzed using both parametric and nonparametric person-fit statistics. A researcher uses an IRT model because it fits the data or because it has some favorable statistical characteristics. Nonparametric IRT models have the advantage that they more easily fit an empirical dataset than parametric IRT models. Which model should be used to analyze item score patterns? From the results obtained in this study, I tentatively conclude that a parametric IRT model may be preferred to a nonparametric IRT model.

Using different kinds of person-fit statistics with the Rasch model resulted in a higher percentage of item score patterns classified as misfitting compared to Mokken's IRT models and, perhaps more interesting, it resulted in different kinds of misfit. In this

study, \mathcal{P} was too conservative in classifying an item score patterns as misfitting. For $\alpha = 0.05$ no item score patterns were classified as misfitting. Because \mathcal{P} only gives an upper bound to the significance probability I conclude that for a set of items for which the IRFs do not intersect a parametric approach may be preferred to a nonparametric approach. As Junker (2001) noted, "the Rasch model is a very well-behaved exponential family model with immediately understandable sufficient statistics (...). Much of the work on monotone homogeneity models and their cousins is directed at understanding just how generally these understandable, but no longer formally sufficient, statistics yield sensible inferences about examinees and items". If I interpret "sensible inferences" here as "inferences about misfitting response behavior" it seems more difficult to draw these inferences using nonparametric IRT than using parametric IRT. This finding suggests that future research should be aimed at improving the power of nonparametric person-fit statistics, and specifying under which conditions these methods can be used as alternatives to parametric person-fit statistics.

The researcher must also decide which logistic IRT model to use. Person-fit statistics detect small numbers of outliers that do not fit an IRT model. When the Rasch model shows a reasonable fit to the data, this model can detect person-fit because powerful statistical tests are available for this model compared to other IRT models (Meijer & Sijtsma, 2001). Moreover, I could distinguish different types of misfit using different kind of statistics. Although I used data from an intelligence test, the person-fit statistics illustrated in this paper could also be applied to data from personality testing to detect aberrant response patterns related to different test taking behaviors.

References

- Agresti, A. (1996). *An introduction to categorical data analysis*. New York: Wiley.
- Dragow, F., Levine M. V., & McLaughlin, M. E. (1991). Appropriateness measurement for some multidimensional test batteries. *Applied Psychological Measurement, 15*, 171-191.
- Dragow, 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.

Drenth, P. J. D. (1969). *Test voor niet Verbaal Abstractie (Test for Nonverbal Abstract Reasoning)*. Lisse: Swets & Zeitlinger.

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

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

Glas, C. A. W. (1988). *Contributions to testing and estimating Rasch models*. Unpublished doctoral dissertation, University of Twente, the Netherlands.

Glas, C. A. W., & Ellis, J. L. (1993). *Rasch scaling program*. Groningen: iec ProGAMMA, the Netherlands.

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. Clausen (Eds.), *Measurement & prediction*. Princeton: Princeton University Press.

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

Junker, B. (2001). On the interplay between nonparametric and parametric IRT, with some thoughts about the future. In: A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.). *Essays on item response theory*. New-York: Springer-Verlag.

Junker, B. W., & Sijtsma, K. (2001). Nonparametric item response theory [special issue]. *Applied Psychological Measurement*, 25 (3).

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

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

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*. New York: Chapman & Hall.

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

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

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

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

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

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*, 107-135.

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

Meijer, R. R., Sijtsma, K. & Smid, N. G. (1990). An empirical and theoretical comparison of the Rasch and the Mokken approach to IRT. *Applied Psychological Measurement*, *14*, 283-298.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. The Hague: Mouton.

Mokken, R. J. (1997). Nonparametric models for dichotomous responses. In: W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory*. NY: Springer Verlag.

Mokken, R. J. & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement*, *6*, 417-430

Molenaar, I. W. (1997). Nonparametric models for polytomous responses. In: W. J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory*. NY: Springer Verlag.

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

Molenaar, I. W. & Hoijtink, H. (1995). Person-fit and the Rasch model, with an application to knowledge of logical quantors. *Applied Measurement in Education*, *9*, 27-45.

Molenaar, I. W. & Sijtsma, K. (2000). *MSP5 for windows, a program for Mokken scale analysis for polytomous items*. Groningen: iec ProGAMMA, the Netherlands.

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

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

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research..

Reise, S. P. (2000). Using multivel logistic regression to evaluate person fit in IRT models. *Multivariate Behavioral Research*, 35, 545-570.

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

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

Rosenbaum, P. R. (1987). Probability inequalities for latent scales. *British Journal of Mathematical and Statistical Psychology*, 40, 157-168.

Sijtsma, K. (1998). Methodology review: Nonparametric IRT approaches to the analysis of dichotomous item scores. *Applied Psychological Measurement*, 22, 3-32.

Sijtsma, K. & Meijer, R. R. (2001). The person response function as a tool in person-fit research. *Psychometrika*, 66, 191-207.

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.

Smith, R. M. (1994). A comparison of the power of Rasch total and between-item fit statistics to detect measurement disturbances. *Educational and Psychological Measurement*, 54, 42-55.

Snijders, T. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika*, 66, 331-342.

Stout, W. F. (1990). A new item response theory modeling approach with applications

to unidimensional assessment and ability estimation. *Psychometrika*, 52, 293-325

S-Plus (2000). *Programmer's guide and software*. Data analysis products division, mathsoft, Seattle, WA.

Thissen, D., & Steinberg, L. (1988). Data analysis using item response theory. *Psychological Bulletin*, 104, 385-395.

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

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

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

Wright, B. D., & Panchapakesan, N. (1969). A procedure for sample free item analysis. *Educational and Psychological Measurement*, 29, 23-24.

Zickar M. J., & Robie, C. (1999). Modeling faking good on personality items: An item level analysis. *Journal of Applied Psychology*, 84, 551-563.

Appendix

For each pair of subtest scores a conservative bound based on the hypergeometric distribution can be calculated for the probability that a person has at most $X_{+e} = x_{+e}$ 1s on the easiest subtest. Let J denote the number of items in the two subtests, let J_d denote the number of items in the most difficult subtest and let J_e denote the number of items in the easiest subtest. The cumulative hypergeometric probability has to be calculated bearing in mind that if $X_+ > J_d$ the minimum possible value of X_{+e} is $X_+ - J_d$. Thus

$$P = P(X_{+e} \leq x_{+e} | J, J_e, X_+) = \sum_{w=\max(0, X_+ - J_d)}^{x_{+e}} P(X_{+e} = w | J, J_e, X_+)$$

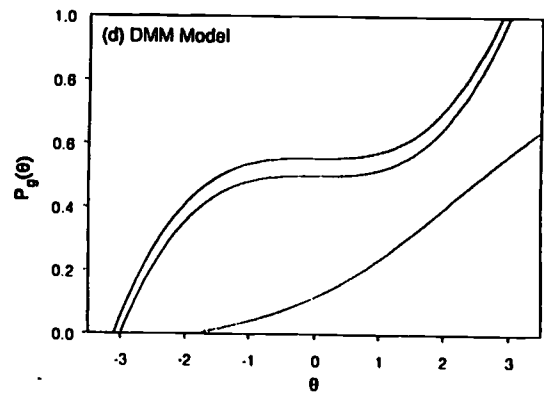
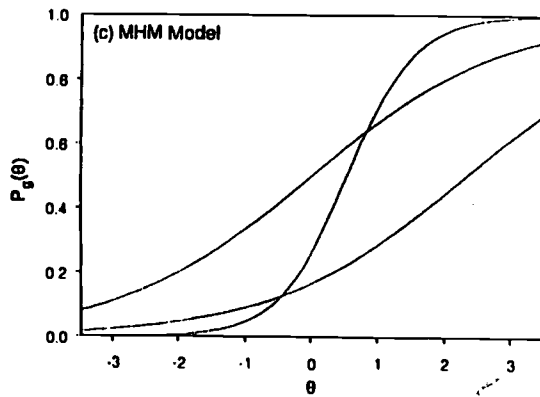
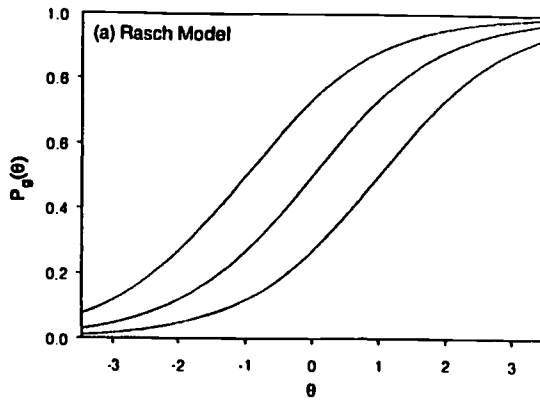
To illustrate this, assume that I have the score pattern (1110100000) and that the items are ordered according to increasing item difficulty, thus the first 5 items in the item ordering are the easiest items and the second 5 items are the most difficult items. Suppose that I would like to compare the subtest score on the first 5 easiest items ($x_{+e} = 4$) with the subtest score on the second 5 most difficult items. Then

$$P = P(X_{+e} \leq 4 | J = 10, J_e = 5, X_+ = 4) = \sum_{w=0}^4 P(X_{+e} = w | J = 10, J_e = 5, X_+ = 4)$$

which equals $P = \frac{\binom{5}{0}\binom{5}{4}}{\binom{10}{4}} + \frac{\binom{5}{1}\binom{5}{3}}{\binom{10}{4}} + \frac{\binom{5}{2}\binom{5}{2}}{\binom{10}{4}} + \frac{\binom{5}{3}\binom{5}{1}}{\binom{10}{4}} + \frac{\binom{5}{4}\binom{5}{0}}{\binom{10}{4}} = 0.0238 + 0.2381 + 0.4762 + 0.2381 + 0.0238 = 1$ However, if we consider the item score pattern (0000010111) then the cumulative hypergeometric probability equals $\frac{\binom{5}{0}\binom{5}{4}}{\binom{10}{4}} = 0.0238$.

Figure Captions

Figure 1. Examples of IRFs for different IRT models.



BEST COPY AVAILABLE

Table 1. M -values for different item score patterns

<i>pattern</i>	$M = -\sum b_i X_i$					
1	1	1	0	0	0	3
2	1	0	1	0	0	2
3	1	0	0	1	0	1
4	0	1	1	0	1	1
5	1	0	0	0	1	0
6	0	1	0	1	0	0
7	0	1	0	0	1	-1
8	0	0	1	1	0	-1
9	0	0	1	0	1	-2
10	0	0	0	1	1	-3

Table 2

Sensitivity of person-fit statistics to different types of deviant behavior, a "+" denotes sensitive and a "-" denotes insensitive to a particular type of deviant behavior

	Misunderstand ing of instruction	Item disclosure	Random response behavior
M	-	+	+
X^2_{dif}	-	+	+
X^2_{ord}	+	-	+
\mathcal{P}	-	+	-

Table 3
 Detection rates for different types of deviant
 response behavior

	Misunderstand ing of instruction	Item disclosure	Random response behavior
<i>M</i>	.071	.453	.764
X^2_{dif}	.153	.872	.663
X^2_{ord}	.871	.231	.343
\mathcal{P}	.001	.453	.001

Table 4
Detection rates for different person-fit statistics

α	M	X^2_{dif}	X^2_{ord}	\mathcal{P}
0.01	0.027	0.026	0.027	0
0.05	0.070	0.093	0.089	0
0.10	0.125	0.149	0.147	0

Table 5

Significance probabilities for the four most deviant item score patterns according to X^2_{dif} with corresponding M values (items ordered according to increasing item difficulty; a superscript denotes the number of consecutive 0s or 1s thus 1^20^3 denotes 11000)

Person #		X_+	X^2_{dif}	M
754	$010^210^{10}10^2101^20^21010^21010^41^20^2$	12	0.0000	0.0001
285	$0101^301^30^21010^3101^40^31011^40^6$	20	0.0003	0.0013
841	$1^301^{16}0^{20}$	19	0.0004	0.4521
669	$1^401^601^2010^31^201^{12}0^310^2$	27	0.0006	0.0062

Table 6

Significance probabilities for item score patterns classified as misfitting using X^2_{ord} and fitting using X^2_{dif} (item scores ordered according to the order in the test; a superscript denotes the number of consecutive 0s or 1s thus 1^20^3 denotes 11000)

Person #	Item score pattern	X_+	X^2_{ord}	X_{dif}
5	$1^201^601^601^20^21$	16	0.0015	0.072
827	$1^210^3101^4010$	26	0.0016	0.082
101	$1^{17}01^20^41010^{13}$	21	0.0019	0.071

Table 7

Overlap of the number of item score patterns classified as misfitting at $\alpha=0.05$

	M	X^2_{ord}	X^2_{dif}
M	69		
X^2_{ord}	22	88	
X^2_{dif}	19	30	93

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

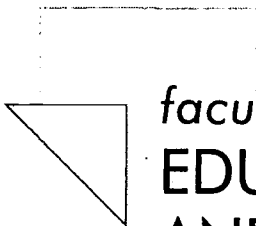
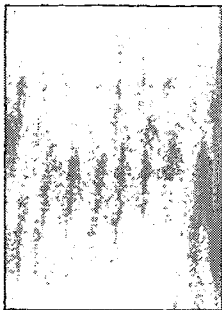
- RR-01-02 R.R. Meijer, *Diagnosing Item Score Patterns using IRT Based Person-Fit Statistics*
- RR-01-01 W.J. van der Linden & H. Chang, *Implementing Content Constraints in Alpha-Stratified Adaptive testing Using a Shadow test Approach*
- RR-00-11 B.P. Veldkamp & W.J. van der Linden, *Multidimensional Adaptive Testing with Constraints on Test Content*
- RR-00-10 W.J. van der Linden, *A Test-Theoretic Approach to Observed-Score equating*
- RR-00-09 W.J. van der Linden & E.M.L.A. van Krimpen-Stoop, *Using Response Times to Detect Aberrant Responses in Computerized Adaptive Testing*
- RR-00-08 L. Chang & W.J. van der Linden & H.J. Vos, *A New Test-Centered Standard-Setting Method Based on Interdependent Evaluation of Item Alternatives*
- RR-00-07 W.J. van der linden, *Optimal Stratification of Item Pools in a-Stratified Computerized Adaptive Testing*
- RR-00-06 C.A.W. Glas & H.J. Vos, *Adaptive Mastery Testing Using a Multidimensional IRT Model and Bayesian Sequential Decision Theory*
- RR-00-05 B.P. Veldkamp, *Modifications of the Branch-and-Bound Algorithm for Application in Constrained Adaptive Testing*
- RR-00-04 B.P. Veldkamp, *Constrained Multidimensional Test Assembly*
- RR-00-03 J.P. Fox & C.A.W. Glas, *Bayesian Modeling of Measurement Error in Predictor Variables using Item Response Theory*
- RR-00-02 J.P. Fox, *Stochastic EM for Estimating the Parameters of a Multilevel IRT Model*
- RR-00-01 E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *Detection of Person Misfit in Computerized Adaptive Tests with Polytomous Items*
- RR-99-08 W.J. van der Linden & J.E. Carlson, *Calculating Balanced Incomplete Block Designs for Educational Assessments*
- RR-99-07 N.D. Verhelst & F. Kaftandjieva, *A Rational Method to Determine Cutoff Scores*
- RR-99-06 G. van Engelenburg, *Statistical Analysis for the Solomon Four-Group Design*
- RR-99-05 E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *CUSUM-Based Person-Fit Statistics for Adaptive Testing*
- RR-99-04 H.J. Vos, *A Minimax Procedure in the Context of Sequential Mastery Testing*

BEST COPY AVAILABLE

- RR-99-03 B.P. Veldkamp & W.J. van der Linden, *Designing Item Pools for Computerized Adaptive Testing*
- RR-99-02 W.J. van der Linden, *Adaptive Testing with Equated Number-Correct Scoring*
- 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 Adaptive Testing 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*

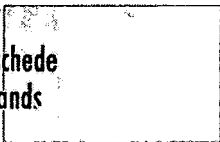
...

Research Reports can be obtained at costs, Faculty of Educational Science and Technology, University of Twente, TO/OMD, 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



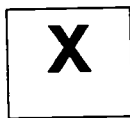


U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)



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").