ED 389 751 TM 024 384

AUTHOR	Meijer, Rob R.; Sijtsma, Klaas
TITLE	Detection of Aberrant Item Score Patterns: A Review of Recent Developments. Research Report 94-8.
INSTITUTION	Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.
PUB DATE	Nov 94
NOTE	28p.
AVAILABLE FROM	Bibliotheek, Faculty of Educational Science and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Foreign Countries; *Identification; *Item Response Theory; *Nonparametric Statistics; Norms; *Scores; *Test Items
IDENTIFIERS	*Aberrance; Item Score Patterns; *Person Fit Measures

ABSTRACT

Methods for detecting item score patterns that are unlikely (aberrant) given that a parametric item response theory (IRT) model gives an adequate description of the data or given the responses of the other persons in the group are discussed. The emphasis here is on the latter group of statistics. These statistics can be applied when a nonparametric model is used to fit the data or when the data are described in the absence of an IRT model. After discussion of the literature on person-fit methods, the use of person-fit statistics in empirical data analysis is briefly discussed. In some situations, the analysis of item score patterns might reveal more information about examinees than the analysis of test scores. Finding an aberrant pattern does not explain the reason for the aberrance. A full person-fit analysis requires additional research into the motives, strategies, and backgrounds of the examinees who deviate from the statistical norm set by the model or group. (Contains 47 references.) (Author/SLD)

Detection of Aberrant Item Score Patterns: A Review of Recent Developments

Research Report 94-8

Rob R. Meijer

Klaas Sijtsma

faculty of
EDUCATIONAL SCIENCE
AND TECHNOLOGY

University of Twente

Department of
Educational Measurement and Data Analysis

Detection of Aberrant Item Score Patterns:
a Review of Recent Developments

Rob R. Meijer
Klaas Sijtsma

Detection of aberrant item score patterns: A review of recent developments, Rob R. Meijer and Klaas Sijtsma - Enschede: University of Twente, Faculty of Educationa: Science and Technology, November, 1994. - 22 pages.

Abstract

Methods for detecting item score patterns that are unlikely (aberrant) given that a parametric item response theory (IRT) model gives an adequate description of the data or given the responses of the other persons in the group are discussed. The emphasis here is on the latter group of statistics. These statistics can be applied when a nonparametric model is used to fit the data or when the data are described in the absence of an IRT model. After the discussion of the literature on person-fit methods, the use of person-fit statistics in empirical data analysis is briefly discussed.

Key words: appropriateness measurement, item response theory, nonparametric item response models, person-fit analysis.

Detection of Aberrant Item Score Patterns:
a Review of Recent Developments

Consider an examinee with a high ability who takes an exam that consists of, say. 30 items. The test contains approximately equal numbers of easy, medium, and difficult items. Assume that the easy items refer to topics from the easy subject matter, and so forth. Thus item difficulty corresponds with difficulty of subject matter. Note that, in general, this is not necessarily true since difficult questions can be asked about easy topics and easy questions about difficult topics. However, for this example our assumptions are appropriate and, in general, they are realistic in many educational tests. Furthermore, suppose that our examinee is anxious to obtain a sufficient result for the exam. To realize this goal he intensely studies the more difficult parts of the subject matter and neglects the easier parts. Let us suppose that this strategy is unusual among the students who prepare for this particular exam.

When administered the test, this examinee will get many of the items of medium and high difficulty correct but many of the easy items incorrect. His response pattern deviates from the patterns produced by examinees of about equal ability who divided their attention more evenly across the subject matter and who thus have response patterns with relatively many 1s (a 1 score corresponds to a correct answer) for easy items, fewer 1s for items of medium difficulty, and the fewest 1s for the most difficult items. Given a test model that assumes that the probability of obtaining the correct answer on any item increases with the ability, our examinee has produced an aberrant pattern, whereas the other students mostly have produced nonnal patterns.

Several other examples can be given of unexpected response patterns. For example, guessing, cheating, or alignment errors may produce aberrant patterns of item scores. Several methods have been proposed to detect persons with item score patterns that are unexpected (aberrant). In this article a review of several of these methods is given. Furthermore, it is briefly discussed how the practitioner may use the methods to detect aberrant item score patterns.

Person-Fit Analysis

By means of an educational test a person's ability or achievement level is measured. Usually, a function of the item scores is used to estimate overall test performance. For example, the number of correct answers may be used for this purpose.

Once information about overall test performance is obtained, additional information for diagnostic purposes may be based on the pattern of item scores. For example, information may be obtained about certain (sub)abilities (Tatsuoka, 1985), cheating or guessing on educational tests (Levine & Rubin, 1979), membership in a subgroup that was initially not identified as relevant for the investigation, for example, a subgroup suffering from a language deficiency (Van der Flier, 1982), and scoring and other clerical errors (Hulin, Drasgow, & Parsons, 1983, chap. 4).

In item response theory (IRT) several models have been proposed in which the probability of obtaining a particular item score is explained by characteristics of a person (the latent ability) and characteristics of the items (e.g., the item difficulty) (e.g., Hambleton & Swaminathan, 1985). IRT models are formulated so as to permit the derivation of consequences that can be checked empirically.

Parametric and nonparametric IRT models can be distinguished. In parametric IRT models (e.g., Birnbaum, 1968; Rasch, 1960), the probability of a correct item response is a parametrically defined function of the person and item parameters. For the purpose of parameter estimation sometimes the distribution of the person parameter is parametrically specified as well. In nonparametric IRT models (e.g., Holland, 1981; Mokken, 1971; Stout, 1990) the probability of a correct item response is defined as an ordinal function of the latent ability. Besides, the exact form of the distribution of the person parameter is left free.

Several differences exist between nonparametric and parametric IRT models. For example, nonparametric models are generally less restrictive than parametric models (Meijer, Sijtsma, & Smid, 1990; Sijtsma & Verweij, 1992). As a consequence, nonparametric models often require more items in a test than parametric models. However, nonparametric models lead to measurement on an ordinal scale, whereas parametric models lead to measurement on a interval or ratio scale. Furthermore, parametric models allow the evaluation of measurement precision as a function of the latent ability by means of the information function (Lord, 1980, p. 21). As in classical test theory, nonparametric models determine accuracy of measurement uniformly for the whole population. Methods for the investigation of the empirical fit of parametric and nonparametric IRT models are discussed by, for example, Meijer et al. (1990) and Sijtsma and Verweij (1992).

If a model fits the data, it may be investigated in a new sample whether persons exist with an item score pattern that is very unlikely given the model. In the 1980s much research has been done to construct methods for detecting such persons. Information about aberrance can be used in addition to the estimate of the overall test performance. This kind of rese:.rch is known as appropriateness measurement research or, more often, person-fit research (Weiss, 1983, chap. 5).

Until today aberrant response behavior has predominantly been investigated in the context of ability and achievement measurement. As a consequence, most person-fit research has concentrated on dichotomously scored items. For person-fit research in the context of polytomous items refer to Drasgow, Levine, Williams, McLaughlin, and Candell (1989).

Two approaches can be distinguished in person-fit research. In the first approach the likelihood of an item score pattern is evaluated given that a parametric IRT model fits the data. If the likelihood of a score pattern is small, this pattern is classified as aberrant. Examples of this approach can be found in Levine and Drasgow (1982), Drasgow, Levine, and Williams (1985), Drasgow, Levine. and McLaughlin (1987), Molenaar and Hoijtink (1990), and Drasgow, Levine, and McLaughlin (1991). In the second approach an item score pattern is evaluated given the item score patterns of the other persons in the group or given that a nonparametric model fits the data. In general, item score patterns with many 1s on items on which most persons in the group have a 0 score and vice versa are classified as aberrant. Examples of this approach have been presented by Van der Flier (1980, 1982), Harnisch and Linn (1981), Harnisch (1983), and Meijer (1994).

More recently, most person-fit research has concentrated on the first approach. If a nonparametric IRT model is used to describe the data, or if no IRT model is used at all, the researcher has to rely on methods from the second approach.

Parametric Person-Fit Analysis

Two groups of statistics can be distinguished. In the first group, statistics are based on a residual which reflects the difference between the item scores expected according to the model, and the observed item scores. Examples of these statistics are provided by Wright (1977), Wright and Stone (1979, pp. 165-190), Wright and Masters (1982), and Tatsuoka (1984). In the second group, statistics are based on a likelihood function. Most parametric person-fit research uses statistics from the second group. Let θ denote the latent ability, and let $\hat{\theta}$ denote the maximum likelihood estimator of θ . Given that the IRT model under consideration fits the data, it is assumed that $\hat{\theta}$ is not a good measure of a person's ability if the likelihood of the item score pattern given $\hat{\theta}$ is small in comparison with the likelihoods of the other patterns with the same estimated ability. In the context of the three-parameter logistic model, Levine and Rubin (1979) proposed the first likelihood statistic denoted l. Drasgow et al. (1985) found that l is confounded with the ability level. To obtain a better person-fit statistic they proposed a standardized version of l, denoted l_z , that corrects l for its expected value and its variance across independent replications. Furthermore, they showed that l_z is approximately standard normally distributed given that the three-parameter logistic model fits the data.

Molenaar and Hoijtink (1990) used a simplified version of l, denoted M, in the context of the Rasch model. Their research concentrated on the question when to label a person as aberrant given that the Rasch model fits the data. For a given item score vector $\mathbf{X} = \mathbf{x}$, the so-called probability of exceedance was determined as a sum based on all item score patterns with the same number-correct score that have a smaller probability than **X** = **x** or equal probability. This sum can be obtained by calculating the exact probabilities using the Rasch model, or by using a chi-square approximation; see Molenaar and Hoijtink (1990) for the conditions when to use either one of these approaches.

Several studies were conducted in which the power of a statistic to detect aberrant item score patterns was examined (Drasgow, 1982; Drasgow et al., 1985, 1987; Klauer & Rettig, 1990; Kogut, 1987; Levine & Drasgow, 1982; Levine & Rubin, 1979; Liou, 1993; Reise & Due, 1991; Schmitt, Cortina, & Whitney, 1993). For example, Drasgow et al. (1987) investigated the power of l_z to detect item score patterns that were the result of cheating. They found that l_z only provided a high rate of detection for aberrant persons with low or high ability.

Other studies concentrated on the test and the sample characteristics that influence the power of the statistics. For example, Reise and Due (1991) used simulated sample data to study the influence of test length, spread of the item difficulties, and the degree of aberrance (defined by the discrepancy between the discrimination parameters for the aberrant and normal groups in the context of the three-parameter logistic model) on the power of l_z. Holding constant all other factors, an increase in either the test length or the spread of the item difficulties yielded, in general, an increase in the power of l_z. Furthermore, it was found that as the difference between the discrimination parameters of the items used to generate item scores for the aberrant simulees and the discrimination parameters of the items used to generate item scores for the normal simulees increased, the power of l_z increased.

Kogut (1987) showed that the power of *M* was reduced if the item difficulty was estimated in a calibration sample that includes aberrant persons. Furthermore, he found that the amount of the reduction of the power of a statistic varied for different kinds of aberrant response behavior.

In recent years, much theoretical research has been conducted in the context of the three-parameter logistic model (e.g., Drasgow et al., 1985) and the Rasch model (Kogut, 1987; Molenaar & Hoijtink, 1990). As a result, the practitioner is able to determine how improbable an item score pattern is, given that the model holds.

Obviously, the use of parametric person-fit statistics is restricted to studies that use parametric IRT models. In addition, person-fit statistics have been proposed in the context of nonparametric IRT and also outside the context of IRT. These person-fit statistics are discussed in the next section.

Nonparametric Person-Fit Analysis and Group-Based Statistics

Nonparametric person-fit analysis

In nonparametric person-fit research an observed pattern of *k* item scores is aberrant if it is improbable given a nonparametric IRT model. In the context of a nonparametric IRT approach that was also discussed by Mokken (1971) and Mokken and Lewis (1982), Van der Flier (1980, 1982) developed the person-fit statistic *U3*. Another approach was pursued by Rosenbaum (1987). According to Van der Flier, an item score pattern is aberrant if the probability of a specific item score pattern conditional on the number-correct score is small compared to the conditional probability of other patterns with the same number-correct score.

To decide whether an observed pattern has an unusually small probability given a realization of the number-correct score, $X = r$, the probability of exceedance is determined for given $\mathbf{X} = \mathbf{x}$ as a sum based on all item score patterns with the same number-correct score r that have smaller probability of occurrence than $\mathbf{X} = \mathbf{x}$ or equal probability. Since the number of possible item score patterns given $X = r$ equals $\binom{k}{r}$, for tests of realistic length (say, at least 15 items) this may lead to an enormous amount of calculation for certain score groups. To avoid this, the statistic *U3* was developed for which the probability of exceedance can be approximated using the decreasing order of the items according to the proportion correct score on the items (which is the well-known *p*-value). By means of a simulation study Van der Flier (1982) showed that for sets of 17 and 29 items with item proportion-correct values that were either uniformly or normally distributed, the *U3* distributions within different score groups were highly comparable. In an empirical study, it was shown that two groups with different ethnic background could be distinguished by means of *U3*.

Meijer, Molenaar, and Sijtsma (in press) studied the power of *U3* under varying test and person characteristics. It was argued that person-fit analysis in the context of nonparametric IRT modeling is affected by the reliability of the items (a substitute for discrimination power in a nonparametric framework; refer to Meijer, Sijtsma, & Molenaar, 1994; Meredith, 1965), the test length, the kind of aberrant response behavior, and the percentage of aberrant persons in the group. They found that the percentage of aberrant simulees detected increased with increasing mean item reliability, increasing test length, and an increasing ratio of aberrant to normal persons in the group. Besides, persons cheating on the most difficult items in a test were more easily detected than persons guessing blindly on all items. Finally, it was argued that relatively short tests of at least 17 items can be used for person-fit analysis if the items are sufficiently reliable [that is, $\rho \approx .3$; this corresponds to discrimination power approximately equal to 2 and $\theta \sim N(0,1)$].

Meijer et al. (in press) also addressed the reliability or replicability of results obtained by means of *U3* across two independent replications of a test. The percentages of replicable valid aberrants increased with an increase in the mean item reliability, the test length, and the proportion of aberrants in the sample. Obviously, these percentages were smaller than percentages obtained in one replication because percentages on the basis of two repetitions can not exceed the smallest of the two percentages on the basis of separate samples. For realistic item reliabilities and test lengths, the mean percentages of replicable valid aberrants across pairs of independent replications were often 10 to 25% smaller than corresponding mean percentages based on single samples from the same population.

Group-Based Statistics

Attempts to analyze item score patterns outside the context 2of IRT are often based on the work of Guttman (1950) and Sato (1975). Given that the items in the sample are ordered according to increasing difficulty (decreasing proportion-correct score) group-based person-fit statistics are, in general, based on the number of 1s to the right of every 0 in a particular item score vector **X** = **x**. This count gives the number of Guttman errors on the basis of all item pairs among *k* items. An item score pattern that does not contain Guttman errors is a Guttman vector, whereas an item score pattern with the maximum number of Guttman errors (all 1s to the right of all 0s) is a reversed Guttman vector. Item scores are often weighted with the item totals (i.e., the numbers of correct answers per item) or the item proportions correct to determine the degree of deviance of an item score pattern.

Sato (1975) proposed the caution index *C*. By means of *C* the observed item score vector of an individual is compared with the vector containing the item total frequencies. $C = 0$ if an observed item score vector is a Guttman vector; $C = 1$ if the covariance of an observed item score pattern with the vector containing the item total frequencies equals zero; and $C > 1$ if this covariance is negative. High positive values of *C* indicate that the item score vector may be unlikely given the overall ordering of the items on the basis of the item total frequencies. Because *C* does not have a fixed upper bound, the interpretation of its values may be problematic (Harnisch & Linn, 1981). Therefore, Harnisch and Linn (1981) proposed the modified caution index C^*. The values of C^* range from 0 to 1; $C^* = 0$ if the observed item score vector is a Guttman vector and $C^* = 1$ if the observed item score vector is a reversed Guttman vector.

Tatsuoka and Tatsuoka (1982) proposed the norm conformity index *NCI*. *NCI* compares the number of conformal pairs of item scores (i.e., the pairs of item scores with a 1 for the easier item and a 0 for the more difficult item) with the number of Guttman errors. $NCI = 1$ if **X** is a reversed Guttman vector and $NCI = -1$ if **X** is a Guttman vector.

Tatsuoka and Tatsuoka (1983) investigated the power of NCI to detect aberrant persons on a set of arithmetic items. They compared two groups of students. One group was far off the mastery level and made many different kinds of errors. The other group was close to the mastery level and only made sophisticated errors. The item difficulty ordering was different for the two groups.

It was shown that students who only made sophisticated errors and were included in the group far off the mastery stage were classified as aberrants.

whereas the inclusion of these same students in the group which was close to mastery resulted in their classification as normal. This empirical example illustrated that *NCI* obtains a relatively high positive value (indicating aberrance) if the item score pattern deviates from the majority of patterns.

In a study in which the power of several group-based statistics was compared, Harnisch and Linn (1981) discussed three other simple group-based indices. The agreement index (*A*) equals the sum of the weighted item scores in which the weights are the proportions correct of the items. The disagreement index (*D*) is obtained by subtracting this weighted sum from the maximum value it can attain given the number-correct score [*A*(max)]. The dependability index (*E*) is obtained by dividing *A* by *A*(max). Both *D* and *E* were proposed to reduce the confounding of *A* with the number-correct score.

To classify a pattern as aberrant, critical values have been proposed for most nonparametric and group-based person-fit statistics. For example, Sato (1975) suggested that for *C* a value higher than .5 indicates aberrance. Harnisch and Linn (1981) considered item score patterns for which C^* was higher than .3 as aberrant. These critical values were based on the experience with only one or two empirical data sets. However, for the *U3* statistic Van der Flier (1980, 1982) showed that a standardized version of *U3* is approximately standard normally distributed provided that the items can be identically ordered for each measurement value on the scale.

Studies comparing Nonparametric and Group-based Person-Fit Statistics

Harnisch and Linn (1981) used empirical data from a reading test and from a math test to obtain the correlation between *C*, C^*, *A*, *D*, *E*, and *NCI*, and the correlation of these indices with the number-correct score. Most indices correlated between .66 and .99 with each other, with the exception of *A* which correlated between .13 and .77 with the other statistics. Note that *A* is strongly confounded with the number-correct score. Most indices correlated approximately .5 with the number-correct score on both tests. For *A* these correlations were .99. However, C^* correlated -.02 and -.21 with the number-correct scores on both tests.

Meijer (1994, chap. 4) proposed six person-fit statistics that weight each Guttman error by the reliability of the two items that are involved in a particular error, and the distance between the proportions correct of these items. The idea (Meijer, 1994, p. 73) behind this was that normal persons are expected to make Guttman errors in particular on items that are unreliable and on items with item difficulties that are narrowly spaced, whereas aberrant persons are expected to make additional Guttman errors on reliable items and items that are widely spaced. By means of simulated data the power of these statistics was compared with the power of *U3* and the number of unweighted Guttman errors. The power of the number of unweighted Guttman errors approximately equalled the power of *U3* and the other statistics. Furthermore, Meijer (1994, p. 84) found a correlation varying from -.01 to -.28 between all statistics and the number-correct score.

These results shed an interesting light on the results obtained by Harnisch and Linn (1981). On the basis of two empirical data sets they preferred C^* because it correlated -.02 and -.21 with the number-correct score. For the simulated data sets analyzed by Meijer (1994, chap. 4) the correlations of most statistics with the number-correct score had the same magnitude. Thus, using this correlation criterion to select a person-fit statistic, the count of the number of Guttman errors seems a good and extremely simple alternative for more complex person-fit statistics.

Rudner (1983) used simulated data to compare several person-fit statistics. He used four residual person-fit statistics (Wright & Masters, 1982), the likelihood statistic *l*, and two group-based statistics, *C* and *NCI*. He found correlations between these statistics varying from .12 to .97. To investigate the power of these indices two cases were distinguished. In one case, for a minority of persons several correct responses were randomly selected and changed into incorrect responses thus producing spuriously low number-correct scores. In the other case, several incorrect responses were changed into correct responses thereby producing spuriously high number-correct scores. It was investigated whether the persons with the spuriously high or low scores were correctly classified as aberrant by each statistic. The conclusion was that the power generally increased with the number of altered item scores. It was further concluded that *C* and *NCI* yielded the most stable results. However, none of the statistics had the highest power in all conditions that were investigated.

On the basis of the literature, a systematic comparison of the person-fit statistics with respect to the relationship with the number-correct score and the rate of detection of aberrants seems hardly possible. Most studies are incomplete, and the characteristics of the data sets used are not always clearly described. For example, Harnisch and Linn (1981) decided on the basis of only two empirical data sets that C^* was better suited to detect aberrant item score patterns than other person-fit statistics. Rudner (1983) did not include *U3* and C^* in his study which obviously makes it impossible to compare the power of these statistics with the power of the other statistics. However, the *U3* statistic has proven to be useful under varying conditions in simulation and empirical research (Van der Flier, 1980, 1982; Meijer et al., in press) and, therefore, seems the most effective statistic to be used in person-fit analysis without a parametric IRT model.

Discussion

In our opinion person-fit statistics are a first step to trace persons whose answering behavior or part of it is the result of other characteristics than the latent ability that the test intends to measure. Person-fit statistics have mostly been used in exploratory analyses where it was vaguely known what kind of aberrant behavior underlied the item scores, or if there had been any aberrant behavior at all. If examinees seem to be aberrant on the basis of their item scores, additional information should be collected which might help to understand the causes of unexpected item score patterns.

As a first step in identifying aberrant patterns the researcher might use statistical criteria such as a statistical test (e.g., Drasgow et al., 1985; Molenaar & Hoijtink, 1990; Van der Flier, 1980, 1982) if a parametric or a nonparametric model is used, or a cut score (e.g., Harnisch & Linn, 1981; Meijer, 1991) in a group-based context for assessing person-fit.

Sole reliance on the pattern of item scores and statistical criteria for their assessment might easily lead to wrong conclusions. For example, consider aberrance as a result of cheating or copying. Much power research using simulated data has defined cheating as the result of a less able examinee copying correct answers on difficult items from a more able neighbor. The assumptions are that a cheater has relatively low ability, selects a high ability-examinee to sit next to at the exam, and is only willing to take the risk of being caught if the most difficult items are involved. Another assumption may be that the high ability neighbor always produces correct answers on these most difficult items.

The typical item score pattern of a cheater would thus show more correct answers on difficult items than expected on the basis of his low ability.

Although the assumptions seem reasonably effective in power research using simulated data, their practical validity could be questioned. For example, someone who copies extensively from the examinee he or she happens to sit next to will not appear aberrant, unless the person copied from was also in this category. Another example is a weak examinee who copies the answers on the most difficult items from his neighbor, but whether he would appear aberrant depends also on the ability of the neighbor copied from. In other words, cheating on exams may manifest itself in several forms and may lead to many different types of item score patterns, some appearing aberrant and others normal. If the researcher suspects that for some persons item scores are due to cheating, he may use one of the statistics particularly devised to detect cheating that were discussed by Frary (1993): refer also to Frary, Tideman, and Watts (1977). This may reduce the number of false negatives.

False positives could arise, for example, from (1) attributing a pattern with relatively many ones on the difficult items to cheating while the mechanism producing it had been guessing (assuming that the items are multiple choice); (2) a learning strategy that stressed the more difficult parts of the subject matter and neglected the easier parts; or (3) a simple clerical error made by the instructor while scoring the exam. It may be noted that, in general, several different causes might lead to the same kind of pattern. Together with the many faces of particular aberrant behaviors such as cheating this further underlines the need for collecting additional information about examinees and items that should be used to understand the causes of aberrant item score patterns. As noted by Molenaar and Hoijtink (1990), blindly removing persons from the data because they have improbable item score patterns should be avoided.

Conclusion

In contrast to the assumptions of most IRT and other test models, the behavior of some examinees while solving items from tests is driven by several abilities or traits rather than one. Examinees may also differ with respect to the strategies used to solve the items. As a result, the test performance of such examinees often can not be adequately explained by means of a single test score. Different solution strategies, whatever they are, might be reflected in different patterns of item scores more obviously than in different test scores. Thus the analysis of item score patterns might reveal more information about examinees than the analysis of test scores. In this review several statistics have been discussed that can be used to detect aberrant item score patterns.

Finding an aberrant pattern does not provide the explanation for this aberrance. The application of person-fit analysis techniques thus might easily lead to the detection of aberrant patterns whereas the reasons for this aberrance are poorly understood. Therefore, a full person-fit analysis requires additional research into the motives, the strategies, and the background of those examinees that deviate from the statistical norm set by the model or the group.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick, *Statistical theories of mental test scores*. Reading: Addison-Wesley.

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

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

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

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*, 67-86.

Drasgow, F., Levine, M. V., Williams, E. A., McLaughlin, M. E., & Candell, G. L. (1989). Modeling incorrect responses to multiple-choice items with multilinear formula score theory. *Applied Psychological Measurement, 13*, 285-299.

Frary, R. B. (1993). Statistical detection of multiple-choice 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.

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 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.

Holland, P. W. (1981). When are item response models consistent with observed data? *Psychometrika, 46,* 79-92.

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

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

Kogut, J. (1987). *Detecting aberrant response patterns in the Rasch model.* (Report 87-2). Enschede: University of Twente, Department of Education.

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

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

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

Meijer, R. R. (1991). Het detecteren en interpreteren van afwijkende responsepatronen op een studietoets [Detection and interpretation of aberrant response patterns on an ability test]. In J. Hoogstraten & W. J. van der Linden (Eds.), *Methodologie Onderwijsresearchdagen '91.* Amsterdam: Stichting Kohnstammfonds voor Onderwijsresearch.

Meijer R. R. (1994). *Nonparametric person fit analysis.* Unpublished doctoral dissertation. Amsterdam: Vrije Universiteit.

Meijer, R. R., Molenaar, I. W., & Sijtsma, K. (in press). Influence of person and group characteristics on nonparametric appropriateness measurement. *Applied Psychological Measurement.*

Meijer, R. R., Sijtsma, K., & Molenaar, I. W. (1994). *Reliability estimation for single dichotomous items based on Mokken's IRT model.* Manuscript submitted for publication.

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

Meredith, W. (1965). Some results based on a general stochastic model for mental tests. *Psychometrika, 30,* 419-440.

Mokken, R. J. (1971). *A theory and procedure of scale analysis.* New York/Berlin: De Gruyter.

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

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Nielsen & Lydiche.

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

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

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

Sato, T. (1975). *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.

Sijtsma, K., & Verweij, A. C. (1992). Mokken scale Analysis: theoretical considerations and an application to transitivity tasks. *Applied Measurement in Education, 5,* 355-373.

Stout, W. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 55,* 293-325.

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

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classification approach. *Journal of Educational Statistics, 10*, 55-73.

Tatsuoka, K. K., & Tatsuoka, M. M. (1982). Detection of aberrant response patterns and their effect on dimensionality. *Journal of Educational Statistics, 20*, 221-230.

Tatsuoka, K. K., & Tatsuoka , M. M. (1983). Spotting erroneous rules of operation by the individual consistency index. *Journal of Educational Measurement, 7*, 215-231.

Van der Flier, H. (1980). *Vergelijkbaarheid van individuele testprestaties [Comparability 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.

Weiss, D. J. (1983). *New horizons in testing*. New York: Academic Press.

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

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

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

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

RR-94-8 R.R. Meijer & K. Sijtsma, *Detection of aberrant item score patterns: A review of recent developments*

RR-94-7 W.J. van der Linden & R.M. Luecht, *An optimization model for test assembly to match observed-score distributions*

RR-94-6 W.J.J. Veerkamp & M.P.F. Berger, *Some new item selection criteria for adaptive testing*

RR-94-5 R.R. Meijer, K. Sijtsma & I.W. Molenaar, *Reliability estimation for single dichotomous items*

RR-94-4 M.P.F. Berger & W.J.J. Veerkamp, *A review of selection methods for optimal design*

RR-94-3 W.J. van der Linden, *A conceptual analysis of standard setting in large-scale assessments*

RR-94-2 W.J. van der Linden & H.J. Vos, *A compensatory approach to optimal selection with mastery scores*

RR-94-1 R.R. Meijer, *The influence of the presence of deviant item score patterns on the power of a person-fit statistic*

RR-93-1 P. Westers & H. Kelderman, *Generalizations of the Solution-Error Response-Error Model*

RR-91-1 H. Kelderman, *Computing Maximum Likelihood Estimates of Loglinear Models from Marginal Sums with Special Attention to Loglinear Item Response Theory*

RR-90-8 M.P.F. Berger & D.L. Knol, *On the Assessment of Dimensionality in Multidimensional Item Response Theory Models*

RR-90-7 E. Boekkooi-Timminga, *A Method for Designing IRT-based Item Banks*

RR-90-6 J.J. Adema, *The Construction of Weakly Parallel Tests by Mathematical Programming*

RR-90-5 J.J. Adema, *A Revised Simplex Method for Test Construction Problems*

RR-90-4 J.J. Adema, *Methods and Models for the Construction of Weakly Parallel Tests*

RR-90-2 H. Tobi, *Item Response Theory at subject- and group-level*

RR-90-1 P. Westers & H. Kelderman, *Differential item functioning in multiple choice items*

<u>Research Reports</u> can be obtained at costs from Bibliotheek, Faculty of Educational Science and Technology, University of Twente, 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