Global, local and graphical person-fit analysis using person response functions

Emons, W.H.M.; Sijtsma, K.; Meijer, R.R.

Published in:
Psychological Methods

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

Publication date:
2005

Psychological Methods
2005, Vol. 10, No. 1, 101–119

1082-989X/05/$12.00 DOI: 10.1037/1082-989X.10.1.101

Global, Local, and Graphical Person-Fit Analysis Using Person-Response Functions

Wilco H. M. Emons and Klaas Sijtsma
Tilburg University

Rob R. Meijer
University of Twente

Person-fit statistics test whether the likelihood of a respondent's complete vector of item scores on a test is low given the hypothesized item response theory model. This binary information may be insufficient for diagnosing the cause of a misfitting item-score vector. The authors propose a comprehensive methodology for person-fit analysis in the context of nonparametric item response theory. The methodology (a) includes H. Van der Flier's (1982) global person-fit statistic *U3* to make the binary decision about fit or misfit of a person's item-score vector, (b) uses kernel smoothing (J. O. Ramsay, 1991) to estimate the person-response function for the misfitting item-score vectors, and (c) evaluates unexpected trends in the person-response function using a new local person-fit statistic (W. H. M. Emons, 2003). An empirical data example shows how to use the methodology for practical person-fit analysis.

Keywords: aberrant response patterns, misfitting item-score vectors, nonparametric item response theory, person-fit analysis, person-response function

A long tradition in psychological assessment has argued for investigating the quality of individual score patterns on tests. In one line of research additional information obtained from the arrangement of the scores on different subtests has been used to predict criterion behavior (e.g., Davison & Davenport, 2002). In another line of research the arrangement of individual item scores has been investigated and compared with what has been expected on the basis of a test model. This research has usually been referred to as person-fit research (e.g., Drasgow, Levine, & McLaughlin, 1987; Meijer & Sijtsma, 2001). Person-fit analysis may, for example, lead to the conclusion that John's performance on an intelligence test reflects an unusual lack of concentration on the easiest items instead of his true intelligence level. Likewise, in a personality inventory the person-fit analysis of Mary's performance may indicate an unusual fear of being evaluated, which is greater or stronger than her true level of introversion. Although one hopes that valid tests produce valid results for each individual being tested, the examples show that this may not always be true. Person-fit analysis helps to identify cases of invalid individual test performance and may be helpful to suggest remedies for the problems involved.

Person-fit researchers (e.g., Drasgow et al., 1987; Klauer, 1991; Molenaar & Hoijtink, 1990; Reise, 2000; Reise & Widaman, 1999) have suggested several statistics for identifying misfitting vectors of item scores on the J items from a test; see Meijer and Sijtsma (2001) for a comprehensive review. These person-fit statistics all assume a particular item response theory (IRT) model (e.g., Embretson & Reise, 2000; Sijtsma & Molenaar, 2002) to fit the test data. Person-fit statistics have been used, for example, to identify examinees with inconsistent item-score patterns on items that required similar cognitive skills (Tatsuoka & Tatsuoka, 1983), to investigate the effect of test anxiety on test performance (Birenbaum, 1986), and to detect respondents who faked on a personality test to convey a favorable impression (Zickar & Drasgow, 1996).

By evaluating the whole vector of J item scores simultaneously, person-fit statistics allow the conclusion that a particular IRT model either does or does not fit a respondent's item-score vector. In this sense, most person-fit methods are global methods that identify misfit but do not help to identify the type of behavior that caused the misfit. An exception is due to Klauer (1991; also, see Meijer, 2003),

Wilco H. M. Emons and Klaas Sijtsma, Department of Methodology and Statistics, Tilburg University, Tilburg, the Netherlands; Rob R. Meijer, Department of Research Methodology, Measurement, and Data Analysis, University of Twente, Enschede, the Netherlands.

We are grateful to N. Bleichrodt, W. C. M. Resing, and P. J. D. Drenth for making available the data of the Revised Amsterdam Child Intelligence Test.

Correspondence concerning this article should be addressed to Wilco H. M. Emons, Department of Methodology and Statistics, FSW, Tilburg University, P. O. Box 90153, 5000 LE Tilburg, the Netherlands. E-mail: w.h.m.emons@uvt.nl

who proposed a method that identifies person misfit caused by violations of either unidimensional measurement, item discrimination, or local independence under the Rasch (1960) model. Also, on the basis of work by Wright and Stone (1979) for the Rasch model, Smith (1985) assumed that a test can be divided into nonoverlapping subtests for which large discrepancies between observed and expected item scores indicate person misfit. This approach is flexible in that it allows for groupings of items based not only on difficulty, as is common in person-fit research (Meijer & Sijtsma, 2001), but also on item content or presentation order of the items. However, Type I error rates were found to be highly sensitive to the distributions of ability and the item parameters, and Molenaar and Hoijtink (1990) found that several standardizations of the statistics could not alleviate these deficiencies. Li and Olejnik (1997) found that the sampling distributions of the statistics discussed by Smith (1986) deviated significantly from the standard normal distribution.

In this article, we propose a comprehensive person-fit methodology that gives more insight than does a single statistic into the possible causes of a misfitting item-score vector. Thus, this methodology helps the practitioner to reach a better diagnosis of respondents' misfitting item scores. The methods we use are sensitive to the ordering of the items according to their difficulty. Other orderings may be useful, but are the topic of future research. Another concern in person-fit analysis is that an item-score vector of only J observations is available for each respondent. The number J typically ranges from, say, 10 to 60. This small number of data points makes person-fit analysis hazardous from a statistical point of view. In particular, low power may render misfitting item-score vectors difficult to detect, resulting in detection rates that are too low. Because of limited testing time for each ability to be tested, the lengthening of tests to well over, say, a hundred items, is not a realistic option.

An alternative to both the limited value of a binary outcome (that provides little information for individual diagnosis) and the small number of data points (that provides little power, implying modest detection rates) may be to seek various other sources of information about an item-score vector's misfit. The combination of these sources may lead to a more accurate decision about misfit or fit and also to more insight into the cause of an item-score vector's misfit. This article discusses a methodology for a more comprehensive person-fit analysis that uses various sources of person-fit information. The methodology compensates to some extent for the necessarily small number of data points in person-fit analysis and facilitates the interpretation of misfit. The methodology includes the global person-fit statistic *U3* (Emons, Meijer, & Sijtsma, 2002; Van der Flier, 1982); a new graphical method that uses kernel smoothing to estimate the person-response function (PRF), based on Ramsay's (1991; also see Douglas & Cohen, 2001; Habing, 2001) smooth estimates of item response functions (IRFs); and a new local person-fit statistic (Emons, 2003) that evaluates unexpected trends in the PRF. The context of the research was nonparametric item response theory (NIRT; Junker, 1993; Ramsay, 1991; Sijtsma & Molenaar, 2002; Stout, 1987). An empirical data example shows how to use the methodology in practical person-fit analysis.

In this study, we restricted ourselves to intelligence data (mostly due to space limitations), but person-fit methods are also useful for analyzing personality data. For example, Reise and Waller (1993) explored the study of person fit in personality measurement by analyzing empirical data from the Multidimensional Personality Questionnaire (Tellegen, 1982). They noted that because of measurement error or faulty responding it can be difficult to distinguish persons fitting the particular trait from persons misfitting the trait. To reduce the opportunities for misfit due to measurement error or faulty responding, they used unidimensional subscales and information from detection scales that identify inconsistent answer behavior. A person-fit statistic was effective in identifying persons who were not responding according to a particular IRT model but had not been identified by detection scales.

Methodology for Comprehensive Person-Fit Analysis

Methodology Proposal

We suggest three stages in a comprehensive person-fit analysis. The technical details of the methods used at each stage are discussed below. The first stage entails traditional person-fit analysis, the second and third are new.

Global analysis. Van der Flier's (1982) global person-fit statistic *U3* was used to identify fitting and misfitting item-score vectors.

Graphical analysis. Kernel smoothing is used to estimate the PRFs for the misfitting item-score vectors that were flagged by *U3*. The PRF gives the probability of a correct response (scored 1) as a function of the difficulty of the items. This function is nonincreasing when the J IRFs in a test do not intersect (Sijtsma & Meijer, 2001). For each misfitting item-score vector, the graph of the PRF is inspected for local increases.

Local analysis. Deviations from the monotone nonincreasing trend in the PRFs are tested locally using a statistical test proposed by Emons (2003).

The combination of global testing, graphical inspection of the PRF for misfitting item-score vectors, and local testing of increases found in the PRF together help to better diagnose the misfit indicated by *U3*, but it may be noted that the final diagnosis also depends on other information. For example, knowing that one individual is dyslexic or that

another individual has a history of fearing personal evaluation may be important, and catching a cheating student red-handed overrules any other source of information. As the psychologist usually does not know the cause of an atypical item-score vector, for a better understanding of the potential causes, background information about individual examinees needs to be incorporated into the diagnostic process. Depending on the application, such information may come from previous psychological-ability and achievement testing, school performance (tests and teacher's accounts), personality testing, clinical and health sources (e.g., about dyslexia, learning, and memory problems), and social–economic indicators (e.g., related to language problems at home). Exactly how this background information may be used to explain person-fit statistics and PRFs is the topic of our present ongoing research. In the next subsection, some examples of misfit and the use of the proposed methodology are given.

Examples

Test anxiety. Assume a respondent was presented the items in an intelligence test in order of ascending difficulty and that he or she suffered from test anxiety during, say, the first 10 items in the test (the easiest items) and performed much better on the other more difficult items. Furthermore, assume that the resulting atypical item-score vector was detected by the *U3* statistic. To facilitate the diagnosis of the cause of the misfit, we estimated the PRF (Figure 1A) for this respondent. Given the effect of test anxiety described, the PRF started at a low value for the lower levels of item difficulty, increased for the items of average difficulty when test anxiety has diminished, and decreased when item difficulty increased further. For a respondent of average or high ability and for items that are administered in ascending difficulty ordering, test anxiety typically results in this bell-shaped curve. For a low-ability respondent, however, the PRF probably would look more like a near-horizontal curve located at a narrow range of low-response probabilities. We return to this latter case in the *Item disclosure* section. For the PRF in Figure 1A, a local test statistic (Emons, 2003), to be explained below, may be used to determine whether the increase in the first 10 items is significant. When a significant local test result is found, the researcher may use the bell shape for further diagnostic decision-making, possibly taking additional background information into account.

Item disclosure. When a test is used for selection with important consequences for individuals, people may be tempted to obtain information about the type of test questions or even about correct answers to particular items before they take the test in an attempt to improve their test performance. Item disclosure is a realistic concern because it may result in a larger percentage of correct answers than expected on the basis of the trait being measured. For example, in the Netherlands only a few different types of intelligence tests are available for persons with a higher educational background. Thus, the psychologist has little opportunity to vary the choice of tests and keep test content a secret.

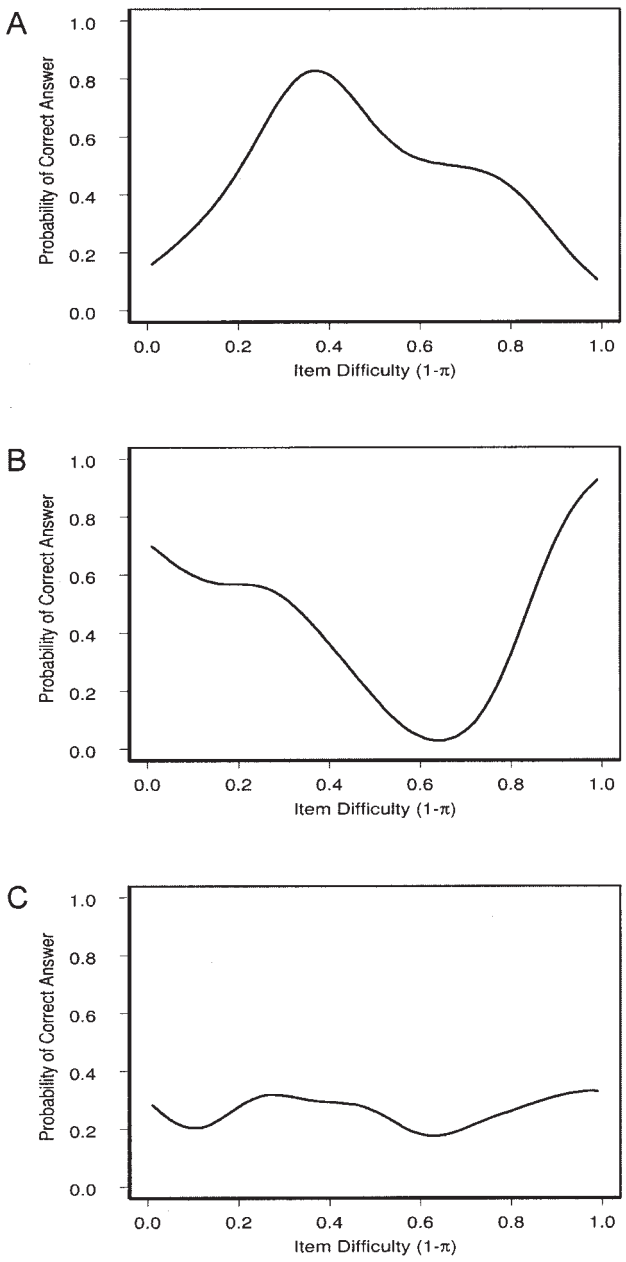

Figure 1. Hypothetical person-response functions for three types of response behavior. A: Test anxiety. B: Item disclosure. C: Random response behavior.

Assume now that a low- or average-ability respondent takes a 50-item intelligence test and tries the 40 relatively easier items but has advance knowledge of the 10 most difficult items (note that the items need not be presented

according to ascending difficulty). Assume that the *U3* person-fit statistic identified the resulting item-score vector as a misfit. A smooth estimate of the PRF shows a decrease for the easiest 40 items because with increasing item difficulty the probability of a correct answer decreases and then shows an increase for the 10 most difficult items because here the respondent gave an unusually high number of correct answers given the item difficulty level; see Figure 1B for this U-shaped PRF. The local test of the PRF may be used to investigate whether the increase in the last 10 items is significant.

Random response behavior. An important line of research in personality assessment has focused on identifying test takers who respond to personality inventories in a random, dishonest, or otherwise deviant fashion. These response sets are threats to the validity of the interpretations made from the resulting profiles (e.g., Grossman, Haywood, & Wasyliw, 1988; Pinsoneault, 2002). One response set that has received considerable attention is that of random response. The random response set includes any approach in which "responses are made without regard to item content" (Graham, 1993, p. 38). Several authors have stressed the particular importance of screening for deviant response sets in criminal populations. Suppose a respondent randomly responds to the four-choice items (say one option is keyed as characteristic of the trait and the others as uncharacteristic) in a personality inventory because he or she is unmotivated to answer the items according to the trait being measured. Assume that the item-score vector that was produced by random response behavior on almost all items was identified by the *U3* statistic. Figure 1C gives a near-horizontal PRF that resulted from an almost constant random response behavior probability of .25 for all J items. This PRF does not deviate noticeably from monotone nonincreasingness, and the local test cannot be applied here. However, given that the items vary widely in difficulty, a near-constant PRF at the random response level for some scale values warns the researcher of aberrant behavior. This example shows the strength of graphical tools for diagnosing aberrant test performance.

Remark about use of other information. A near-horizontal PRF, as in Figure 1C, that is typical of randomly responding cannot be distinguished from a similar PRF that would result from test anxiety for a low-ability respondent or test anxiety for higher ability respondents that resulted from serious panic. Here, other auxiliary information about the respondent may be helpful when evaluating item-score vectors.

For example, suppose that trait-level estimates are available from previous testing (e.g., Drasgow, Levine, & Williams, 1985). Also, assume that a respondent takes different versions of the same test several times per year, for example to measure cognitive improvement after therapy. Given this knowledge, for a high-ability respondent who took the first version of this test, a PRF like that in Figure 1C would probably indicate random response behavior. In this situation, no additional action needs to be taken. However, for a high-stakes test that is taken only once (e.g., for selection purposes), the explanation may be a complete off-day that resulted in panic. Here, one could decide to retest this respondent but under less threatening circumstances. Note that we used the ability level and the test situation (auxiliary information) to make a decision on how to proceed. For a low-ability respondent, a near-horizontal PRF may mean excessive random response behavior due to a test difficulty level that was too high. Here, retesting using a more appropriately tailored test may be reasonable. Auxiliary information based on, for example, the respondent's personal history could indicate, however, that he or she suffered from extreme anxiety. In this case, it would probably not be sufficient to administer an easier test, but perhaps precautions like better instruction and many more exercise items should be taken as well. The use of the ability level is discussed below in an empirical example.

NIRT

Theoretical Introduction to NIRT

The context of this study was NIRT (Sijtsma & Molenaar, 2002). NIRT models assume order restrictions on the IRFs. Let X_j ($j = 1, \ldots, J$) denote the binary random variable for the item responses, with realization $x_j = 1$ for a correct or coded response, and $x_j = 0$ otherwise. Let $X_+ = \Sigma_{j=1}^{J} X_j$ denote the unweighted sum score; let $\hat{\pi}_j$ ($j = 1, \ldots, J$) denote the population proportion of persons with a 1 score on item j; and let $\hat{\pi}_j = N_j/N$ (N is the sample size and N_j the frequency of 1s on item j) be the sample estimate of π_j. We assume that the J items in the test are ordered and numbered from easy to difficult: $\pi_1 \geq \pi_2 \geq \ldots \geq \pi_J$. The probability of obtaining a 1 score is related to the latent trait θ by the IRF: $P_j(\theta) = P(X_j = 1|\theta)$. We assume a scalar θ (this is the unidimensionality assumption of IRT, abbreviated UD). Given UD we assume that item scores are locally independent (assumption LI). A typical NIRT assumption is that the IRFs are monotone nondecreasing in the latent trait (assumption M); that is, for two arbitrary fixed values θ_a and θ_b,

$$P_j(\theta_a) \leq P_j(\theta_b), \text{ whenever } \theta_a < \theta_b; j = 1, \ldots, J.$$

NIRT models that satisfy the assumptions of UD, LI, and M imply that the total score X_+ stochastically orders θ (Grayson, 1988; Hemker, Sijtsma, Molenaar, & Junker, 1997). Stochastic ordering justifies the use of X_+ for ordering persons on θ and is a useful ordering property in practice whenever a test is used to order respondents. Mokken's (1971; also, see Ellis & Van den Wollenberg, 1993; Holland & Rosenbaum, 1986; Junker, 1993) monotone homogeneity model is defined by the assumptions of UD, LI, and M.

For person-fit analysis it is convenient that the IRFs do not intersect, because the same ordering of items by difficulty then applies to each respondent, and this facilitates the interpretation of test performance. Nonintersection for two items i and j means that if we know for a fixed value θ_0 that $P_i(\theta_0) > P_j(\theta_0)$, then

$$P_i(\theta) \geq P_j(\theta), \text{ for all } \theta. \quad (1)$$

This is the assumption of invariant item ordering (IIO; Sijtsma & Junker, 1996). Mokken's model of double monotonicity is defined by the assumptions of UD, LI, M, and IIO. Several methods exist to investigate whether the double monotonicity model fits a set of items (e.g., Hoijtink & Molenaar, 1997; Karabatsos & Sheu, 2004; Mokken, 1971; Sijtsma & Molenaar, 2002). The definitions of the PRF (Sijtsma & Meijer, 2001) and the local person-fit statistic (Emons, 2003), to be discussed shortly, require an IIO.

The Place of the Double Monotonicity Model Within IRT

Figure 2 shows a Venn diagram that explains how the double monotonicity model is related to the monotone homogeneity model and the well known l-, 2-, 3-, and 4-parameter logistic models (abbreviated 1PLM, 2PLM, 3PLM, and 4PLM, respectively). Let δ_j denote the location parameter of the IRF of the 4PLM, α_j the slope parameter, γ_j the lower asymptote, and λ_j the upper asymptote; the 4PLM is then defined as

$$P_j(\theta) = \gamma_j + \frac{(\lambda_j - \gamma_j)\exp[\alpha_j(\theta - \delta_j)]}{1 + \exp[\alpha_j(\theta - \delta_j)]}. \quad (2)$$

The 3PLM is a special case of the 4PLM that assumes that the upper asymptotes are equal to 1 for all J items ($\lambda_j = 1$, $j = 1, \ldots, J$); the 2PLM further narrows the 3PLM by assuming that the lower asymptotes of the IRFs are equal to 0 for all J items ($\gamma_j = 0$, $j = 1, \ldots, J$); and the 1PLM narrows the 2PLM by assuming that the J slopes are equal (normed at $\alpha_j = 1$, $j = 1, \ldots, J$). Thus, the set of tests agreeing with the 1PLM is nested within the set of tests agreeing with the 2PLM, the set of tests agreeing with the 2PLM is nested within the set agreeing with the 3PLM, and the set of tests agreeing with the 3PLM is nested within the set agreeing with the 4PLM. Each of these four models adopts the assumptions of UD, LI, and M, which together define the monotone homogeneity model, and each specifically defines assumption M by adopting a logistic IRF. This means that the 1PLM, the 2PLM, the 3PLM, and the 4PLM are all nested within the monotone homogeneity model (see Figure 2).

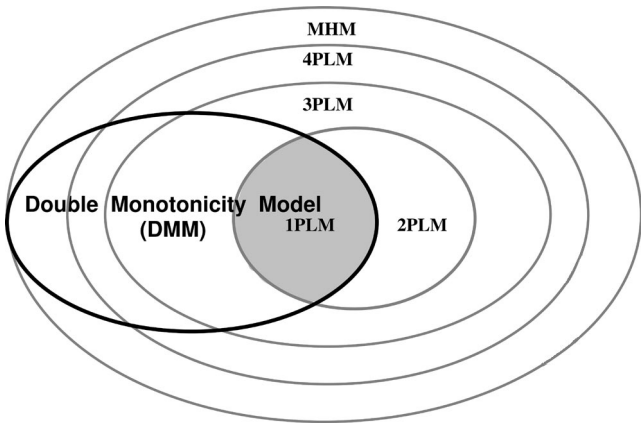

Figure 2. Venn diagram of the relationships between the double monotonicity model and the monotone homogeneity model (MHM), the 4PLM, the 3PLM, the 2PLM, and the 1PLM. PLM = parameter logistic model.

Instead of specifying assumption M by means of logistic IRFs, in the nonparametric context the double monotonicity model assumes that the J IRFs in a test do not intersect (IIO; Equation 1). How does this assumption locate the double monotonicity model in the Venn diagram in Figure 2? First, the double monotonicity model is a special case of the monotone homogeneity model because it is based on the assumptions of UD, LI, and M and, in addition, assumes an IIO. Second, like the double monotonicity model the 1PLM assumes nonintersecting IRFs, but it is more restrictive because the IRFs are logistic curves that are translations of one another along the θ axis. Thus, next to the nested series "monotone homogeneity model–4PLM–3PLM–2PLM–1PLM," Figure 2 also contains the nested series "monotone homogeneity model–double monotonicity model–1PLM." Third, the relationship of the double monotonicity model to the 2PLM, the 3PLM, and the 4PLM is as follows. It is easy to show that IRFs in the 2PLM do not intersect only if their slope parameters α are equal (Sijtsma & Meijer, 2001). Mathematically, the 2PLM has then been reduced to the 1PLM. It follows that there are no IRFs in the 2PLM that are also in the double monotonicity model unless they are also IRFs in the 1PLM. Thus, in Figure 2 the intersection of the sets of the double monotonicity model and the 2PLM is the set of 1PLM items (this is the shaded area). For the 3PLM and the 4PLM the situation is different. Sijtsma and Meijer (2001) showed that if for the 3PLM (1) $\alpha_1 = \alpha_2 = \ldots = \alpha_J$, and (2) $\gamma_1 \geq \gamma_2 \geq \ldots \geq \gamma_J$ and $\delta_1 < \delta_2 < \ldots < \delta_J$, then the J IRFs do not intersect. For the 4PLM, if the conditions 1 and 2 are satisfied and, in addition, (3) $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_J$, then the J IRFs do not intersect. Sets of 3PLM IRFs that satisfy conditions 1 and 2 and sets of 4PLM IRFs that satisfy conditions 1, 2, and 3 also agree with the double monotonicity model. Finally, any sets of monotone IRFs that do not intersect are double monotonicity IRFs. Such IRFs may have lower asymptotes greater than 0, higher asymptotes smaller than 1 (even high-ability examinees have success probability smaller than 1), and multiple in-

flection points (logistic IRFs have one) and may not be symmetric (logistic IRFs are). Figure 3 shows a set of such double monotonicity IRFs.

Figure 2 shows that sets of tests agreeing with the 1PLM also agree with the double monotonicity model and that some tests agreeing with the 3PLM and the 4PLM and some agreeing with the monotone homogeneity model also agree with the double monotonicity model. Thus, the double monotonicity model is more general than the 1PLM and may be seen as a nonparametric version of it.

Desirability of IIO for Person-Fit Analysis

The double monotonicity model based on the IIO assumption is the basis of the person-fit methods used in this study. Do we really need the assumption that the IRFs in the test do not intersect? After all, IRT models that allow the intersection of the IRFs, such as the monotone homogeneity model or perhaps even the 3PLM or the 2PLM, are more likely to fit test data than models based on the IIO assumption (see Figure 2). Below, we argue that person-fit analysis often pursues strong statements about individual test performance at the level of items and that this requires the strong IIO assumption. Without the IIO assumption such statements may be problematic. Next, we argue that, theoretically, IIO is desirable for person-fit analysis in order to have interpretable person-fit results. This is why our methodology is based on the assumption of IIO. Then, we discuss some results from a robustness study, which show that in practical data analysis our methodology is still likely to produce valid results when IIO is not fully satisfied in one's data. The conclusion is that IIO is a desirable property of a person-fit methodology but that in real data analysis small deviations from IIO may be tolerated.

Theoretical discussion of IIO in person-fit analysis. First, we investigate how IRT models that do not have an IIO, such as the 2PLM and the 3PLM, contribute to person-fit analysis. The 2PLM and the 3PLM allow the estimation of an individual's θ from the likelihood based on the vector of all J item scores. If these models do not fit a particular item-score vector, then the respondent's θ estimate, denoted $\hat{\theta}$, may be biased and unduly inaccurate (Drasgow et al., 1985; Meijer, 1997; Schmitt, Chan, Sacco, McFarland, & Jennings, 1999) and, as a result, may not be trusted. This is the kind of information provided by person-fit statistics based on the 2PLM and the 3PLM. It is important for the proper evaluation of an individual's performance on all J items together as summarized in the latent trait θ. Thus, IRT models not implying an IIO are useful for evaluating individual test performance.

For diagnostic purposes, the next question is which item scores caused the person misfit. If all misfitting item-score vectors could be compared with one overall item ordering, this would help greatly to understand misfit at a substantive level. To understand why IIO is needed, suppose that the opposite situation holds, which is that the IRFs intersect as in the 2PLM and the 3PLM. What are the consequences of not having an IIO for the interpretation of individual item-score vectors? As an example, consider IRFs from the 2PLM. Two such IRFs have one intersection point whenever their slope parameters are unequal; and J such items

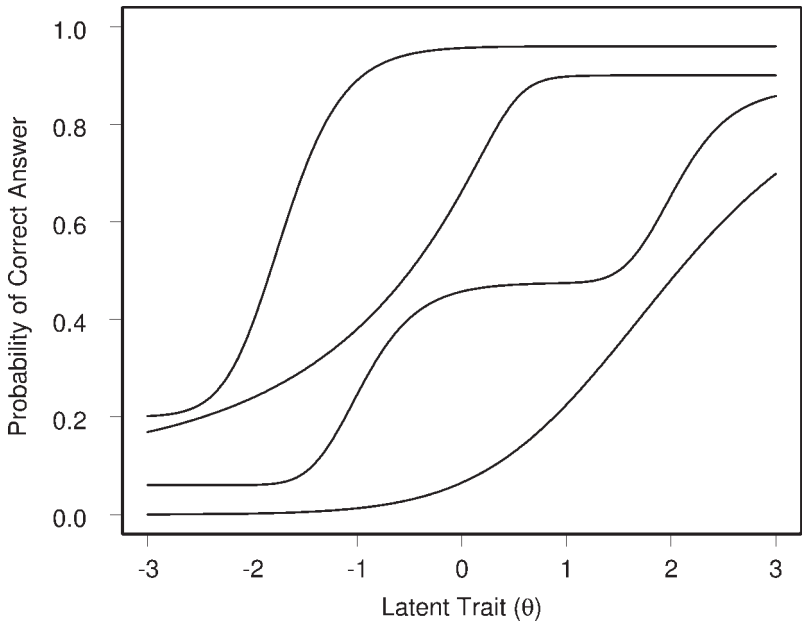

Figure 3. Example of four item response functions satisfying the double monotonicity model.

have $1/2\,J(J-1)$ intersection points defining $1/2\,J(J-1)+1$ disjoint intervals on θ, each characterized by a unique ordering of items by response probability (Sijtsma & Meijer, 2001). For example, Figure 4 shows that four IRFs from the 2PLM define seven disjoint intervals on θ. The figure also shows three θs that have item orderings from easy to difficult: 1–4–3–2 (John), 1–3–4–2 (Mary), and 1–2–3–4 (Cynthia). Note that for Cynthia a 0 score on Item 4 (her most difficult item) and 1 scores on the other three easier items do not produce misfit. However, for John the same item-score vector may produce misfit because for him Item 4 is his second easiest item.

The example shows that under the 2PLM (and also the 3PLM) item ordering depends on the latent trait. Obviously, if item ordering depends on θ (e.g., for $J = 10$, the number of θ intervals is already 46, defining equally many item orderings), an easy interpretation of individual item-score vectors is highly improbable. In the double monotonicity model it is independent of the latent trait due to IIO. IRT models implying IIO (such as the double monotonicity model and the Rasch model) facilitate the interpretation of individual test performance.

Practical discussion of IIO in person-fit analysis. IRT models that have the IIO property in addition to assumption M facilitate the interpretation of individual test results, because each item-score vector can be compared with one overall item ordering, which then serves as a kind of gold standard. IRT models having an IIO are the double monotonicity model and its special case, the 1PLM. Although these are rather restrictive models that sometimes may not fit the data for all J items in a test, there are two reasons why one may be optimistic about the fit of IRT models with an IIO to test data.

First, experienced test constructors often aim to include sets of items that have a wide difficulty range, especially in intelligence and ability testing, and exclude items that have little discriminating power. These two goals together exclude IRFs that are close together and have relatively flat slopes (Figure 5, dotted and dashed IRFs). These would be the IRFs with the highest risk of crossing other IRFs. As a result, the items that are selected in the final test (Figure 5, solid curves) tend to have IRFs that approach the IIO property rather well. Thus, it is likely that data from many real testing applications approximate an IIO because of the way tests are assembled.

Second, for intersecting IRFs that are close together (e.g., Figure 5, solid curves), simulation research (e.g., Sijtsma & Meijer, 2001) has shown that the person-fit methods we used here are robust against departures from IIO. Sijtsma and Meijer (2001) investigated detection rates of aberrant item-score vectors for moderately long tests ($J = 40$) and long tests ($J = 80$) under the 2PLM with slope parameters ranging from 0.8 to 1.2 and under a more general IRT model allowing both lower IRF asymptotes greater than 0, upper

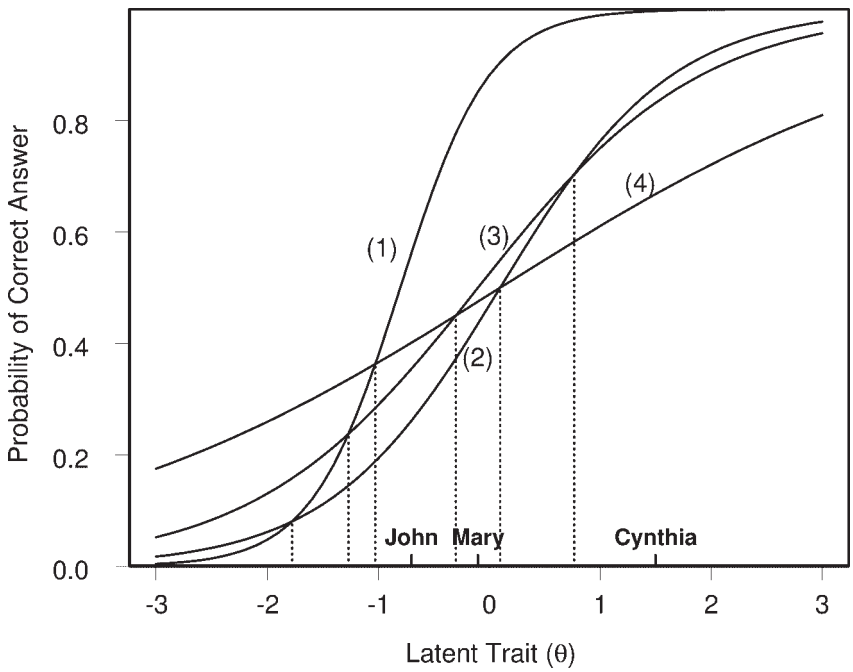

Figure 4. Item response functions from the two-parameter logistic model. Item parameter values are as follows: $\alpha_1 = 2.5$, $\alpha_2 = 1.3$, $\alpha_3 = 1.0$, $\alpha_4 = 0.5$, $\delta_1 = -0.8$, $\delta_2 = 0.1$, $\delta_3 = -0.1$, and $\delta_4 = 0.1$. John = item ordering of 1–4–3–2; Mary = item ordering of 1–3–4–2; Cynthia = item ordering of 1–2–3–4.

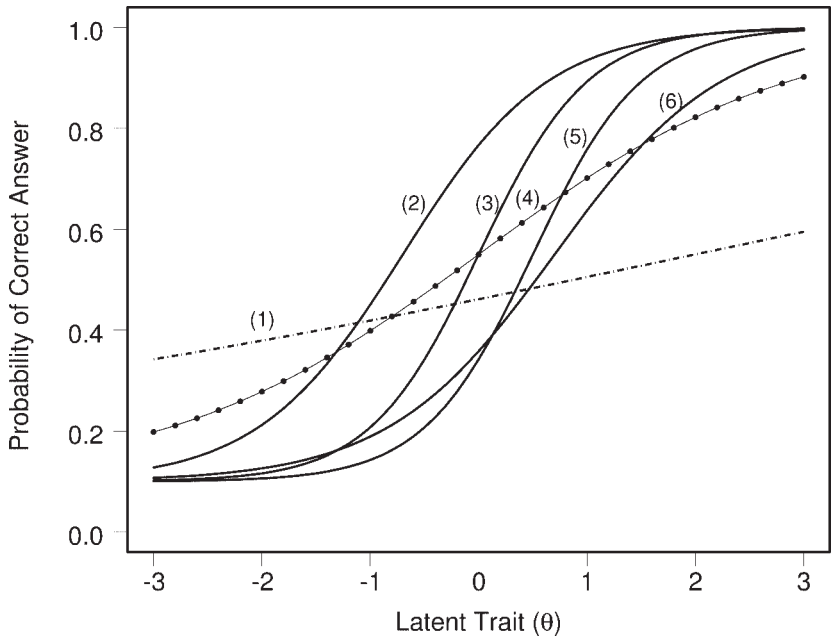

Figure 5. Example of four item response functions with medium discrimination (solid lines) and two item response functions with low discrimination (dotted and dashed lines). Item parameter values are: $\alpha_1 = 0.2$, $\alpha_2 = 1.5$, $\alpha_3 = 2.0$, $\alpha_4 = 0.7$, $\alpha_5 = 2.0$, $\alpha_6 = 1.3$, $\delta_1 = 2.0$, $\delta_2 = -0.7$, $\delta_3 = 0.0$, $\delta_4 = 0.0$, $\delta_5 = 0.5$, and $\delta_6 = 0.7$.

asymptotes smaller than 1, and slopes ranging from 0.8 to 1.2. This choice of slopes created many intersections of the IRFs within a test. At different sides of the intersection point the ordering of success probabilities was opposite, but because the IRFs' slopes were similar, for a fixed θ value success probabilities for different items were close (i.e., the IRFs were rather close). This was designated a mild violation of IIO. It was found that compared with IRFs that were highly comparable but had the same slopes (i.e., IIO held) the detection rates were almost the same. These results indicate that in practical person-fit analysis we can use these person-fit methods even when IIO is not satisfied completely.

Global Analysis: Van der Flier's *U3* Statistic

Let $\mathbf{X} = (X_1, \ldots, X_J)$ denote the vector of J item-score random variables, and let item score vector $\mathbf{x} = (x_1, \ldots, x_J)$ denote the realization of $\mathbf{X}$. Given that items are ordered by decreasing $\hat{\pi}_j$ values, an item-score vector $\mathbf{x}$ with 1s in the first x_+ positions and 0s elsewhere is called a Guttman vector, and a vector with 1s in the last x_+ positions and 0s elsewhere is a reversed Guttman vector. The *U3* statistic (Emons et al., 2002; Meijer, Molenaar, & Sijtsma, 1994; Van der Flier, 1980, 1982) for observed item-score vector $\mathbf{X}$, denoted $U3(\mathbf{X})$, is defined as

$$U3(\mathbf{X}) = \frac{\sum_{j=1}^{X_+} \log\left(\frac{\hat{\pi}_j}{1-\hat{\pi}_j}\right) - \sum_{j=1}^{J} X_j \log\left(\frac{\hat{\pi}_j}{1-\hat{\pi}_j}\right)}{\sum_{j=1}^{X_+} \log\left(\frac{\hat{\pi}_j}{1-\hat{\pi}_j}\right) - \sum_{j=J-X_++1}^{J} \log\left(\frac{\hat{\pi}_j}{1-\hat{\pi}_j}\right)}. \quad (3)$$

For fixed X_+ all terms are constant, except

$$W(\mathbf{X}) = \sum_{j=1}^{J} X_j \log\left(\frac{\hat{\pi}_j}{1-\hat{\pi}_j}\right), \quad (4)$$

which is a random variable and also a function of the random vector $\mathbf{X}$. Equation 3 shows that $U3 = 0$ only if the respondent's item score vector is a Guttman vector, and that $U3 = 1$ only if the respondent's item score vector is a reversed Guttman vector.

Using the sampling theory derived by Van der Flier (1980, 1982) for *U3*, Emons et al. (2002) found that the Type I error rate did not always match the nominal significance level. However, because a higher *U3* corresponds to a less likely item-score vector, the descriptive use of *U3* may involve selecting the highest, say, 5% of the *U3* values to identify atypical item-score vectors. If subsequent research suggests that many of these item-score vectors happen to be aberrant, 5% may have been too low and a higher percentage may be selected. For a distribution in which most of the *U3* values are low, the highest 5% of *U3* values

may contain only a few item-score vectors that are really atypical. Then, only *U3* values may be selected that appear as outliers in the right tail of the empirical *U3* distribution. The effect is that fewer than 5% of the item-score vectors are subjected to further investigation. In a study using simulated data, Karabatsos (2003) found *U3* to be among the 4 best-performing person-fit statistics out of 36 statistics. Future research may replace *U3* by each of the other three statistics in the first stage of our methodology. However, the flexible use of *U3* for selecting possibly atypical item-score vectors as proposed here is likely to make it an effective statistic.

Graphical Analysis

The Person-Response Function

Sijtsma and Meijer (2001) defined the PRF for respondent v as the probability of a correct answer to items measuring θ as a function of their item difficulty. This is formalized by a random variable S_{vj} that takes value 1 if respondent v answered item j correctly and 0 if the answer was incorrect. Let $G(\theta)$ be the cumulative θ distribution. Item difficulty is defined as

$$1 - \pi_j = \int_\theta [1 - P_j(\theta)]dG(\theta), j = 1, \ldots, J, \quad (5)$$

and sample estimates $(1 - \hat{\pi}_j)$ can be used to estimate the ordering of the items. In the context of person-fit analysis, to prevent biased estimates, ideally, the sample should not contain many misfitting item-score vectors (e.g., Meijer & Sijtsma, 2001). In practice, such data may not be available, and the researcher should then be cautious in interpreting his or her results. Under IIO, the item difficulties, $1 - \pi_j$ $(j = 1, \ldots, J)$, theoretically are reverse ordered relative to the response probabilities, $P_j(\theta)$, $j = 1, \ldots, J$. The probability for respondent v to give correct answers as a function of item difficulty, $1 - \pi$, can be written as

$$P_v(1 - \pi) = P(S = 1|1 - \pi, \theta_v). \quad (6)$$

This conditional probability is defined on the continuous scale $(1 - \pi)$ with domain [0,1]. The PRF, $P_v(1 - \pi)$, is nonincreasing under NIRT models that have IIO (Sijtsma & Meijer, 2001). Kernel smoothing (e.g., Fox, 1997; Ramsay, 1991; Simonoff, 1996) was used to obtain a (quasi-)continuous estimate of the PRF. This estimate is convenient for the localization and the interpretation of misfit.

Kernel Smoothed Estimates of the PRF

Kernel smoothing is a nonparametric regression technique (e.g., see Fox, 1997; also, Simonoff, 1996). The input to the method are the J items in the test, which are ordered along the abscissa on the basis of their estimated item difficulties, $1 - \hat{\pi}_j$ (because of IIO, the same item ordering holds for each respondent), and for each respondent the input is his or her 0/1 scores on the J items, which are displayed on the ordinate. Basically, kernel smoothing fits a smooth, nonlinear curve through the 0/1 scores of respondent v as a function of the item difficulties. The result is an estimated PRF. A program for estimating continuous PRFs and variability bands can be obtained from Wilco H. M. Emons.

More specifically, kernel smoothing takes a focal observation indexed 0, here an item difficulty, say, $1 - \hat{\pi}_{j(0)}$ and several of its neighbor item difficulties, and then estimates $P_v(1 - \hat{\pi}_{j(0)})$ as the weighted mean of the item score $x_{vj(0)}$ and the x_{vj}'s of the neighbor items. Weights are assigned by the kernel function, $K(\cdot)$. A subset of observations that is used for estimating one function value is called a window. Each observation $1 - \hat{\pi}_j$ $(j = 1, \ldots, J)$ is the focal point once, and moving to the next focal point means that the left-most item from the previous window does not move along to the new window while the next-difficult item enters the new window from the right. Windows for items at or near the endpoints of the item ordering contain less data. Special precautions take care of the resulting inaccuracy in estimation (e.g., Habing, 2001).

The bandwidth determines the number of observations used in the estimation of the function values. A broader bandwidth means that adjacent estimated function values are more alike because the windows used for estimation are almost identical. Thus, the PRF is estimated relatively accurately (i.e., with little variance), but interesting details may get lost (i.e., this may induce much bias). A narrower bandwidth has the opposite effect: Function values are different because subsequent windows contain few observations, as observations quickly enter and exit the windows as one moves along the item difficulty range. Particular jags in the PRF are visible (and are estimated with little bias), but statistical accuracy is small (i.e., estimates are highly variable). Thus, for a particular application the choice of the bandwidth involves finding the balance between bias and inaccuracy. This is explained in more detail shortly.

Let $z_j = [(1 - \hat{\pi}_j) - (1 - \hat{\pi}_{j(0)})]/h = (\hat{\pi}_{j(0)} - \hat{\pi}_j)/h$, where h is the bandwidth to be defined shortly, and let $K(z_j)$ be the kernel function. The nonparametric regression function we use is defined as

$$\hat{P}_v(1 - \hat{\pi}_{j(0)}) = \frac{\sum_{j=1}^{J} K(z_j)x_{vj}}{\sum_{j=1}^{J} K(z_j)}. \quad (7)$$

For the kernel function we use the standard normal density,

$$K(z_j) = \frac{1}{\sqrt{2 \times 3.141}} \exp^{-z_j^2/2}, \quad (8)$$

which is a common choice. When the standard normal

kernel function is used, each window in fact uses all J observations, but observations further away from the focal observation receive small weights, and truncation eliminates the influence of distant observations. For calculations similar to those performed here, for both several simulated data sets and several real data sets, Emons, Sijtsma, and Meijer (2004) tried bandwidth values $h = 0.05$, 0.09, and 0.13. For $h = 0.05$, they found that PRF estimates are too inaccurate, which lead to many Type I errors; that is, random increases are erroneously taken for signs of real aberrant behavior. For $h = 0.13$, most of the sampling variation was smoothed away and the PRF estimates tended to become linear (except in the tails). Bandwidth $h = 0.09$ tended to show enough detail with sufficient accuracy. It was concluded that each application requires some trial and error to find the best compromise. The PRFs in Figure 1 were estimated using this kernel-smoothing procedure.

The PRF and Local Person Fit

Discrete PRF Estimate

For local person-fit testing, we used a discrete estimate of the PRF (Trabin & Weiss, 1983; also, see Nering & Meijer, 1998; Sijtsma & Meijer, 2001). This discrete estimate may be seen as an extreme version of kernel smoothing, with uniform kernels that do not overlap. First, the J items are ordered by increasing $(1 - \pi)$ values. Then, they are divided into K ordered disjoint subsets, denoted A_k, with $k = 1, \ldots, K$. For simplicity's sake (but not by necessity), each subset contains m items, such that $A_1 = \{X_1, \ldots, X_m\}, A_2 = \{X_{m+1}, \ldots, X_{2m}\}, \ldots, A_K = \{X_{J-m+1}, \ldots, X_J\}$. For respondent v, the expected proportion of correct answers to the items in A_k equals $\tau_{vk} = m^{-1}\sum_{j\in A_k} P_j(\theta_v)$. Given an IIO, an ordering of the items according to the $(1 - \pi_j)$s implies that for each respondent v,

$$m^{-1}\sum_{j\in A_k} P_j(\theta_v) \geq m^{-1}\sum_{j\in A_{k+1}} P_j(\theta_v),$$

$$\text{for all } \theta; \text{ and } v = 1, \ldots, N. \quad (9)$$

For the K item subsets it follows that

$$\tau_{v1} \geq \tau_{v2} \geq \ldots \geq \tau_{vK}, \text{ for all } \theta. \quad (10)$$

Let X_{vj} denote the score of person v on item j. The ordering in Equation 10 is estimated using sample fractions

$$\hat{\tau}_{vk} = m^{-1}\sum_{j\in A_k} X_{vj},\ k = 1,\ldots, K. \quad (11)$$

Figure 6 shows a solid PRF that is decreasing and, thus, in agreement with an IIO (Equation 10). The dashed PRF shows that the proportions correct for the two most difficult item subsets are greater than those of several easier item subsets. This violates Equation 10.

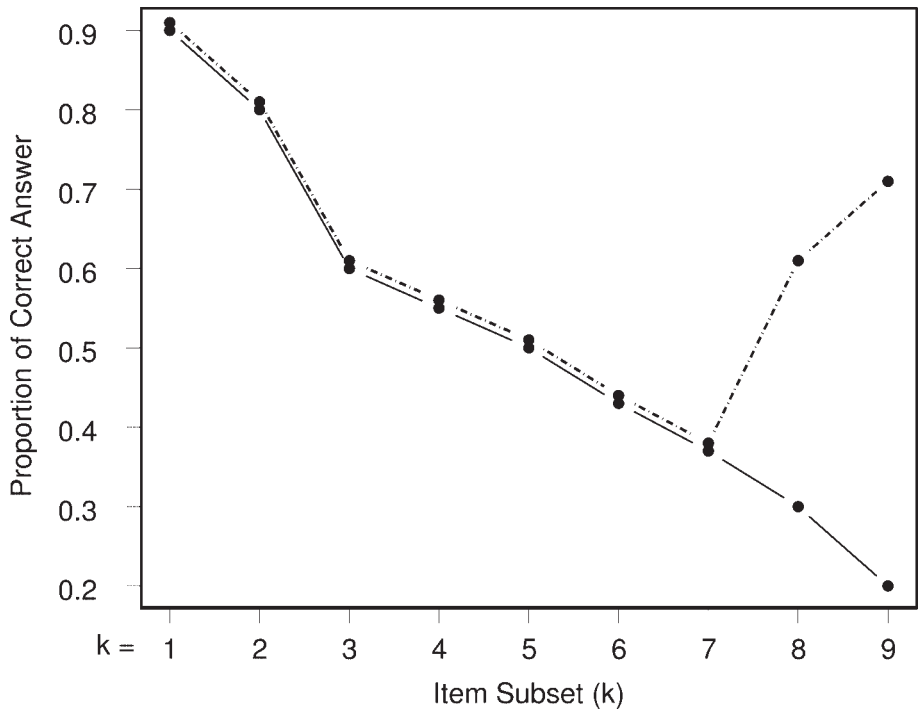

Figure 6. Example of a discrete person-response function indicating expected response behavior (solid line) and a person-response function indicating aberrant response behavior (dashed line).

Testing Local Person Fit

We propose a person-fit statistic that, given an IIO, quantifies the result that in any item subset the correct answers are most likely to be given to the relatively easy items. Define any item vector $\mathbf{Y}$ (e.g., combine subsets A_k and A_{k+1} into one set) in which all J_Y items are ordered by ascending difficulty. Then, count the number of item pairs in $\mathbf{Y}$ in which the easiest item is answered incorrectly while the more difficult item is answered correctly. This is the number of Guttman errors (see, e.g., Meijer, 1994). For respondent v the number of (0,1) patterns on all possible item pairs (including pairs that contain the same item twice) equals

$$G_v = \sum_{j=1}^{J_Y} \sum_{i=1}^{j} (1 - Y_{vj}) Y_{vi}. \quad (12)$$

Person misfit in $\mathbf{Y}$ is revealed by an exceptionally high G value given the expected G value under the postulated NIRT model. For sum score $Y_+ = \sum Y_j$ and realization y_+, we evaluate the probability $P(G \geq g|y_+, J_Y)$ using a theorem proven by Rosenbaum (1987). The theorem says, essentially, that given that the IRFs have IIO (Equation 1), the number of Guttman errors cannot exceed the corresponding number expected under the exchangeable distribution; that is, the number of Guttman errors expected when the response probabilities, $P_j(\theta)$ ($j = 1, \ldots, J_Y$), are equal for all items. This means that the IRFs coincide completely. Because under an NIRT model we cannot evaluate $P(G \geq g|y_+, J_Y)$ directly, we compare it to the corresponding probability under the exchangeable distribution. The probability under the exchangeable distribution is at least as great as the probability of interest under the NIRT model and, thus, provides an upper bound for the probability under the NIRT model. A program to test the local fit can also be obtained from Wilco H. M. Emons.

How is statistic G distributed under the exchangeable distribution? Emons (2003) showed that G is a linear function of the sum of ranks. Thus, under the exchangeable distribution, $P(G \geq g|y_+, J_Y)$ can be obtained from the Wilcoxon's rank-sum distribution. This probability provides an upper bound for $P(G \geq g|y_+, J_Y)$ under IIO. For item subsets containing fewer than 20 items, tables provided by Sprent (1993, p. 319) may be used to obtain probabilities of exceedance. For item subsets containing at least 20 items, G is approximately normally distributed (Sprent, 1993, pp. 116–117). Emons (2003) concluded from a simulation study that for many tests the Type I error rate of G often ranged from 0.02 to 0.03 (nominal $\alpha = .05$), with slightly better results for higher θs. This was found for item sets both with and without an IIO.

Empirical Examples

Amsterdam Revised Child Intelligence Test (RAKIT)

In this section, we used data ($N = 1{,}641$) of the RAKIT (Bleichrodt, Drenth, Zaal, & Resing, 1984; Bleichrodt, Resing, Drenth, & Zaal, 1987) to illustrate the person-fit methodology. The RAKIT measures the cognitive development of children ranging from age 4 to age 11. We analyzed data from four subscales measuring perceptual reasoning: Figure Recognition ($J = 50$), Exclusion ($J = 50$), Quantity ($J = 65$), and Hidden Figures ($J = 45$). For each of the four subscales, the fit of Mokken's (1971) double monotonicity model to the data was investigated using the computer program Mokken Scale analysis for Polytomous items (Molenaar & Sijtsma, 2000). Two results are of main interest here.

First, coefficient H^T (Sijtsma & Meijer, 1992) was used to investigate the IIO assumption for the whole set of J IRFs (the global IIO investigation). According to Sijtsma and Meijer (1992), increasing values of H^T between 0.30 and 1.00 (maximum) mean that the evidence for IIO is more convincing, whereas values below 0.30 indicate important violations of IIO. For the four subsets it was found that $H^T = 0.74$ (Figure Recognition), $H^T = 0.69$ (Exclusion), $H^T = 0.68$ (Quantity), and $H^T = 0.60$ (Hidden Figures). Additional analysis showed no significant intersections between pairs of IRFs (local IIO investigation). Thus, the fit results showed that each subscale well approximates the property of an IIO.

Second, the IRFs had steep slopes relative to the θ distribution; that is, the discrimination power of each item was sufficiently high to have good measurement quality: $H_j \geq 0.36$ for all items from each subscale (as a rule of thumb, $H_j < 0.30$ leads to the rejection of an item; see Sijtsma & Molenaar, 2002, p. 60). This is a favorable property for person-fit analysis. Because the H_js were high, the scalability of the subscales was also high: $H \geq 0.54$ for all four subscales; using Mokken's terminology, these are strong scales ($H \geq 0.50$; Mokken & Lewis, 1982; Sijtsma & Molenaar, 2002, p. 60). The difficulty ordering of the items was estimated from the sample difficulties $1 - \hat{\pi}$. This ordering closely agreed with the administration ordering, from easy to difficult.

Results of the Empirical Person-Fit Analysis

We first summarize the most important results of the global and graphical person-fit analysis for the total sample. Then, we discuss in detail the results of the local person-fit analysis for six individual cases (see Table 1 for the details) who had a *U3* value in the upper 5% range for the 45-item scale Hidden Figures. These cases represent different types of person misfit that were detected using our three-step methodology.

Table 1
Observed Item-Score Vectors from the Subtest Hidden Figures That Are Used for the Six Examples of Graphical and Local Person-Fit Analysis

Case	Observed item-score vector									X_+	*U3*
	1	2	3	4	5	6	7	8	9		
1	00010	11100	11101	11111	10111	01110	00000	00000	00000	19	.28
2	00000	01101	10100	10000	00000	00000	00000	00000	00000	6	.24
3	11011	00000	10101	01111	11100	01000	00000	00000	00000	15	.23
4	11111	10100	11011	11111	11101	11111	01100	00100	10111	32	.35
5	11111	11111	11111	10011	10010	10111	01111	11101	11111	37	.42
6	11111	11111	11111	10111	11110	11110	00011	01111	11101	37	.27

Note. Examples were drawn from a *U3* distribution with $M = .11$, $Mdn = .10$, 25th percentile = .05, and 75th percentile = .15. The cutoff value that was used for identifying misfitting item-score vectors was .23.

Step 1: Global person fit—U3 analysis. Because the subscales had high discrimination, we analyzed global person fit using *U3* as a descriptive statistic. The *U3* frequency distributions in Figure 7 show that each subscale had few extreme *U3* values, which appeared in the right tails of the distributions. For each subscale, we selected the 5% of the item-score vectors with the highest *U3* values and classified them into three X_+ levels, denoted low, medium, and high (not displayed in a figure). Except for Hidden Figures, for the other three subscales more than 70% of the item-score vectors having the highest *U3*s corresponded to the high X_+ level. The subscale Hidden Figures had approximately a uniform distribution of the item-score vectors over the three X_+ levels.

Step 2: Graphical person-fit analysis. For each selected item-score vector, kernel smoothing was used to estimate a (quasi-)continuous PRF. For the subscales Figure Recognition, Exclusion, and Quantity we used a bandwidth $h = 0.08$, and for Hidden Figures we used $h = 0.09$. For each subscale, for low and medium X_+ levels the PRFs had an irregular shape. In particular, for low and medium X_+, some PRFs had a bell shape, such as the example given in Figure 8A. However, most misfitting PRFs for low and medium X_+ showed an increase at medium item difficulty. Examples are given in Figures 8B through 8D. The PRFs for high X_+ levels typically showed a small increase at medium to high item difficulty (see, e.g., right-hand side of Figure 8E). The PRFs for medium X_+ and high X_+ levels rarely showed misfit on the easiest items. Some of the PRFs for high X_+ levels did not show any deviations from the expected nonincreasingness (e.g., Figure 8F). These PRFs result from item-score vectors that contain few incorrect answers that are scattered throughout the test. This pattern may be due to short lapses of concentration or perhaps coincidence. Also, note that when an easy item was failed but several more-difficult items were answered correctly, the failure received much weight and produced a high *U3* but did not affect the shape of the PRF.

Step 3: Local person-fit analysis. Local increases of the PRFs were tested for significance by means of the Wilcoxon's rank-sum test using the number of Guttman errors, G (Equation 12). We illustrate this for the six cases presented in Table 1, which were taken from the 45-item subscale Hidden Figures; see Table 1 for details. For each case, Figure 9 shows the estimated continuous PRF obtained by means of kernel smoothing ($h = 0.09$) and the 90% confidence envelope (evaluated along the ordinate) obtained by means of a jackknife procedure (Emons et al., 2004). The confidence envelopes may be used as a precursor to the Wilcoxon's rank-sum test. This is done as follows.

Consider the null hypothesis, $P(1 - \pi_i) = P(1 - \pi_j)$, which represents the extreme case of no increase between $P(1 - \pi_i)$ and $P(1 - \pi_j)$, and evaluate it against the alternative hypothesis of increase, $P(1 - \pi_i) < P(1 - \pi_j)$. For testing this null hypothesis, assume that the confidence interval for $P(1 - \pi_i)$ was derived from the sampling distribution of $\hat{P}(1 - \hat{\pi}_i)$ under the null hypothesis. If the sample value, $\hat{P}(1 - \hat{\pi}_j)$, is outside the confidence interval for parameter $P(1 - \pi_i)$, it is concluded that the PRF increases significantly between $1 - \pi_i$ and $1 - \pi_j$.

For example, for Case 1 in Figure 9 consider the difficulty values (on the abscissa) approximately equal to .00 and .35 and the corresponding increase in the sample PRF. It is readily verified that the PRF estimate at difficulty value .35 is outside the confidence interval (on the ordinate) for the PRF at difficulty value .00. Thus, the increase is significant. For Case 5, the sample PRF increases between difficulty values of approximately .4 and 1.0. One can verify that the PRF estimate at, for example, difficulty value .6 falls in the confidence region for the PRF at difficulty value .4. This result suggests that the local increase of the PRF between the difficulty values of .4 and .6 is due to sampling error. The PRF estimate at difficulty value 1.0 clearly is outside the confidence interval for the PRF at difficulty value .4. This result suggests a significant increase of the PRF between the difficulty values of .4 and 1.0. These results are corroborated by the Wilcoxon's rank-sum test, to be discussed below.

This procedure demonstrates that the confidence enve-

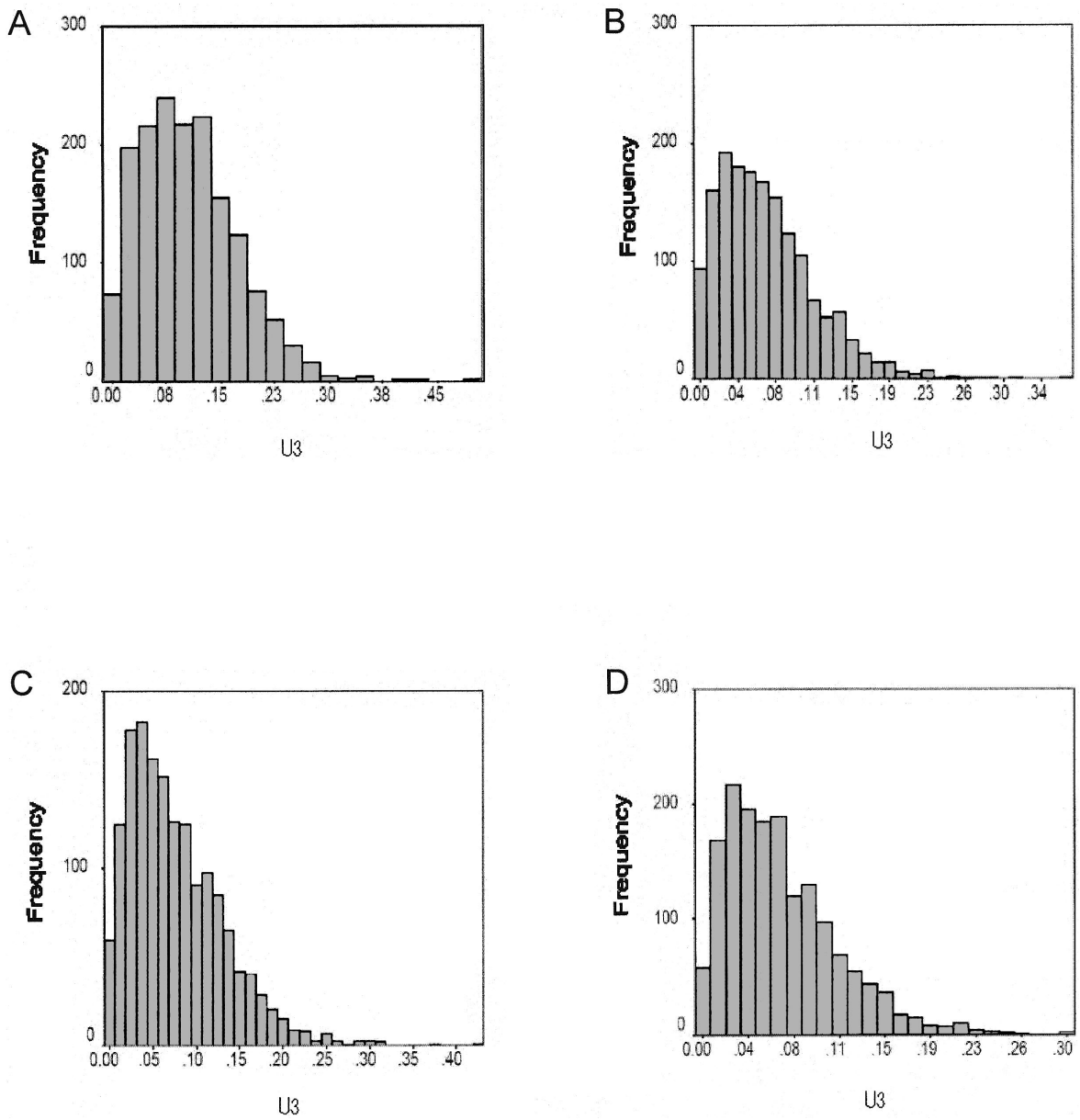

Figure 7. Histograms of *U3* for the four Revised Amsterdam Child Intelligence Test subscales. A: Figure Recognition. B: Exclusion. C: Quantity. D: Hidden Figures.

lopes of the PRFs suggest misfit on the easiest items for Case 1 but not for Case 2; the PRF of Case 3 on the relatively easy items, but not on the easiest items; the PRF of Case 4 on the items of medium and high difficulty; and the PRFs of Cases 5 and 6 on the difficult items. We divided the items into $K = 9$ disjoint subsets, each containing $m = 5$ items; that is, $A_1 = \{X_1, \ldots, X_5\}, \ldots, A_9 = \{X_{41}, \ldots, X_{45}\}$. The discrete approximation of the PRF (see Figure 10) was obtained using Equation 11.

Table 2 gives the results of the local person-fit tests. The item subsets (Table 2, second column) used for local person-fit testing were chosen on the basis of the confidence envelopes (see Figure 9) showing possibly significant local increases of the PRFs. Column 3 shows the number of items in these item subsets. Columns 4, 5, and 7 show the number correct (Y_+), the number of Guttman errors (G), and the significance probability, respectively. The normed number of Guttman errors (G^*) is also presented, and will be discussed below.

For Case 1, the PRF shows a local increase for the first four subsets, A_1 through A_4 (Figure 10A). We combined these subsets into one vector, **Y**, and counted the number of

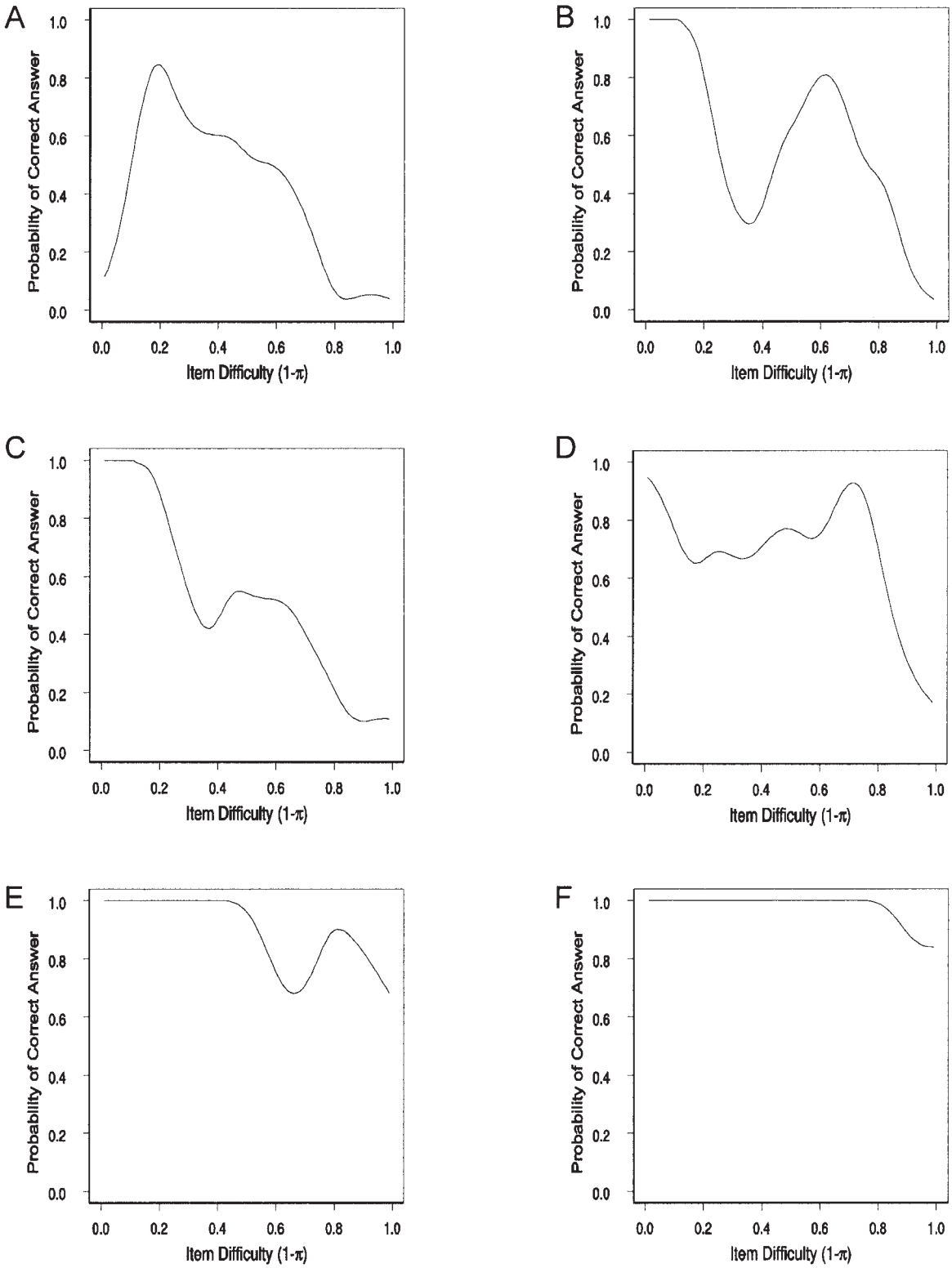

Figure 8. Examples of estimated continuous person-response functions for low latent trait level respondents (A, B, and C), medium latent trait level respondents (D), and high latent trait level respondents (E and F).

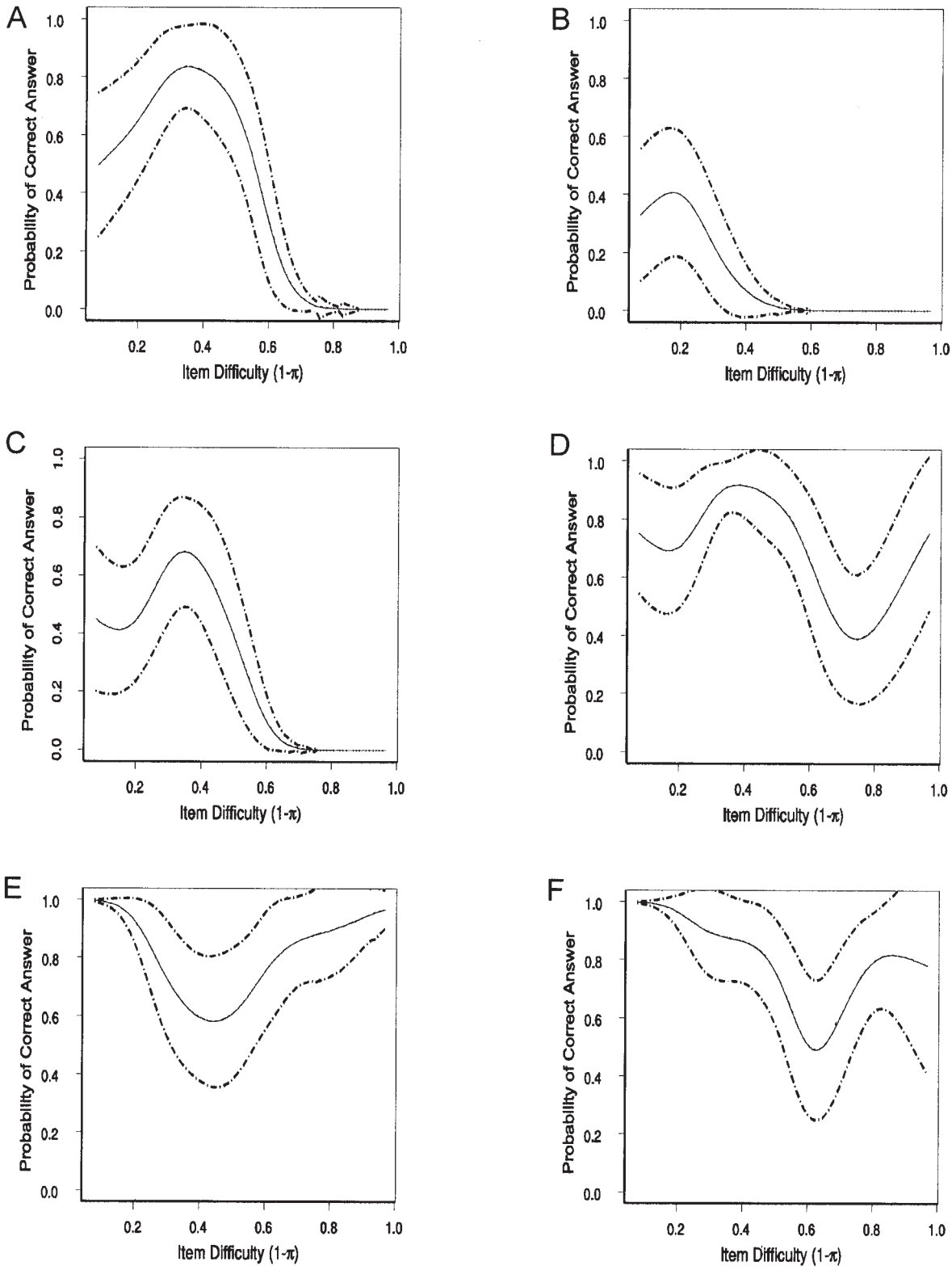

Figure 9. Estimated continuous person-response functions (solid lines) and 90% confidence envelopes (dashed lines) of six cases, subscale Hidden Figures. A: Case 1. B: Case 2. C: Case 3. D: Case 4. E: Case 5. F: Case 6.

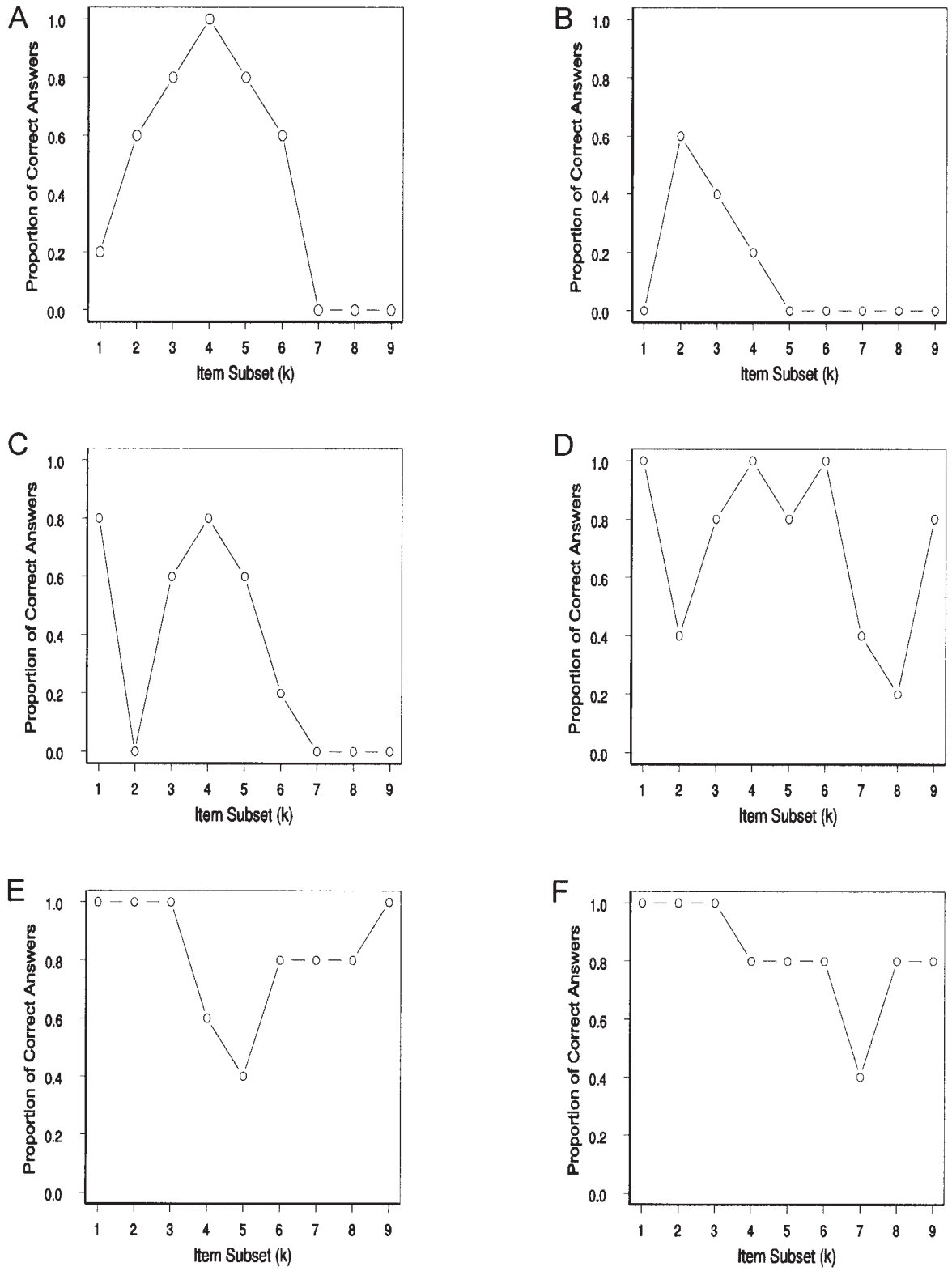

Figure 10. Estimated discrete person-response functions of six cases, subscale Hidden Figures. A: Case 1. B: Case 2. C: Case 3. D: Case 4. E: Case 5. F: Case 6.

Table 2
Results for the Significance Test Using Local Person-Fit Statistic G, for the Six Examples From the Subtest Hidden Figures

Case	Items	J_Y	Y_+	G	G^*	p
1	1–20	20	13	75	.82	.01
2	1–10	10	3	19	.90	.02
3	6–20	15	7	50	.89	.01
4	6–20	15	11	35	.80	.05
	21–30	10	9	6	.67	.40
	36–45	10	5	21	.84	.03
5	21–30	10	6	17	.71	.18
	36–45	10	9	6	.67	.40
	21–45	25	19	88	.77	.03
6	31–40	10	5	22	.88	.02
	31–45	15	10	39	.78	.05

Note. J_Y = number of items; Y_+ = number correct; G = number of Guttman errors; G^* = normed number of Guttman errors.

Guttman errors, G. The upper bound for the significance probability was obtained from the Wilcoxon's rank-sum distribution. For Case 1, $G = 75$, which was significant at the .01 level. The interpretation of G values is enhanced by comparing them with their maximum ($G_{\max}$), given the number of items (J_Y) and the number of correct answers (Y_+). This maximum equals $G_{\max} = Y_+(J_Y - Y_+)$. For Case 1, we have $Y_+ = 13$ given that $J_Y = 20$, so that $G_{\max} = 13 \times (20 - 13) = 91$. The normed number of Guttman errors is $G^* = G/G_{\max}$ (Meijer, 1994; Van der Flier, 1980), which for Case 1 equals 75/91 = .82. This value may be compared with $G^* = 0$, which is the minimum value (characteristic of a Guttman vector), and $G^* = 1$, which is the maximum value (characteristic of a reversed Guttman vector). Another reference value is the expectation of G under the Wilcoxon's rank-sum distribution, which equals $Y_+(J_Y - Y_+)/2$ (e.g., Lindgren, 1993, p. 475). As a result, the expected value of G^* under the Wilcoxon's rank-sum distribution is .5. It follows that G^* values between .5 and 1 indicate that an item-score vector contains more Guttman errors than expected under the null model, whereas values between 0 and .5 indicate fewer Guttman errors. Given the reference values of 0 (minimum), .5 (expectation) and 1 (maximum), we conclude that $G^* = .82$ is high. More information may be available from the empirical distribution of G^* in a group of respondents (cf. Rudner, 1983; Tatsuoka & Tatsuoka, 1982).

A school-behavior inventory (Bleichrodt, Resing, & Zaal, 1993) showed that Case 1 scored low on emotional stability. Furthermore, the current subscale in the RAKIT battery was preceded by the more difficult subscale Learning Names. This may suggest that Case 1 was seriously discouraged by the difficulty of the preceding subscale and, as a result, gave many incorrect answers to the first and easiest items of the current subscale. This is an example of how knowledge of actual school behavior and the difficulty of subscales may help to interpret person-fit results.

For Case 2, the local person-fit test on the items in A_1 and A_2 was significant ($G = 19$, $p < .02$; $G^* = .90$). For Case 3, the test showed significant misfit on the relatively easy items ($G = 50$, $p < .01$; $G^* = .89$). For Case 4, the PRF showed three local increases, which were each tested for significance. A significant result was found for Items 6 through 20 (A_2, A_3, A_4; $G = 35$, $p < .05$; $G^* = .80$) and for Items 36 through 45 (A_8, A_9; $G = 21$, $p < .03$; $G^* = .84$). The local increase for Items 21 through 30 was not significant (A_5, A_6; $G = 6$, $p < .40$; $G^* = .67$). This local increase for the discrete PRF was not shown by the estimated continuous PRF (Figure 9D). Thus, for an appropriately chosen bandwidth kernel smoothing reveals the more persistent deviations and suppresses the unimportant ones. The high *U3* value for Case 4 (see Table 1) can be explained by the misfit for relatively easy items and relatively difficult items. Case 4 had also scored high on a measure of general school performance. The zigzag pattern of correct and incorrect answers for this high-ability respondent may be an indication of test anxiety as an explanation of the observed misfit. In practical applications, this result may motivate further assessment of the respondent's test anxiety. For Case 5, three local tests were done. Increases at Items 21 through 30 and 36 through 45 were not significant, but the increase for Items 21 through 45 was (A_5 through A_9; $G = 88$; $p < .03$; $G^* = .77$). This misfit ranged over 25 items, which may explain the high *U3* value. For Case 6, significant misfit was found for Items 31 through 40 ($G = 22$, $p < .02$; $G^* = .88$) and for Items 31 through 45 ($G = 39$, $p < .05$; $G^* = .78$). Thus, Case 6 had some relatively easy items incorrect, but 8 items correct out of the 10 most difficult items. Because the RAKIT was administered individually, answer copying was no explanation, and the interpretation of this result is not straightforward.

Discussion

The usual person-fit statistics lead to the binary conclusion that an IRT model either fits or does not fit an item-score vector. Graphical analysis of person-response functions followed by testing of local deviations in person-response functions leads to more insight into possible causes of item-score misfit. We used methods from nonparametric IRT because of their flexibility in data analysis. We argue that parametric and nonparametric IRT models based on the assumptions of UD, LI, and M provide person-fit information that is useful to identify respondents whose test scores may not be trusted. In addition, we argue that an invariant item ordering is needed to better understand a misfitting item-score vector flagged by a global person-fit statistic. Even though this is an important restriction on data analysis, many tests may approach an invariant

item ordering because of how test construction is typically done. Also, simulation results (Emons, 2003; Sijtsma & Meijer, 2001) indicated robustness of person-fit methods against violations of invariant item ordering in test data.

The simultaneous use of *U3*, the estimation of PRFs by means of kernel smoothing, and the use of the upper bound for the number of Guttman errors based on the Wilcoxon's rank-sum distribution are new in person-fit research. One of the improvements currently under investigation is the estimation of confidence envelopes using a jackknife procedure (see Figure 9). Such regions may help to better visualize the violations that are candidates to be tested for significance. Also, they may help researchers to better recognize and evaluate trends in person-response functions.

Several artificial and real data examples clarified the use of our methodology. The use of auxiliary information seems to be highly important to reaching good decisions. This is a topic for future research. It is our firm belief that the use of graphical methods in combination with global and local statistical testing, to be expanded in future applications with models that incorporate the use of relevant background information, can make a useful contribution to the practice of psychological and educational diagnosis. More power for finding misfitting item-score vectors may also come from first determining whether the items in a test have the IIO property and then using the proposed person-fit methodology on newly tested individuals. However, in some applications this situation may be too idealistic, because researchers may also want to investigate misfit for the sample used to calibrate the test.

References

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

Bleichrodt, N., Drenth, P. J. D., Zaal, J. N., & Resing, W. C. M. (1984). *Revisie Amsterdamse Kinder Intelligentie Test* [Revision of Amsterdam Child Intelligence Test]. Lisse, the Netherlands: Swets & Zeitlinger.

Bleichrodt, N., Resing, W. C. M., Drenth, P. J. D., & Zaal, J. N. (1987). *Intelligentie-meting bij kinderen* [Intelligence measurement of children]. Lisse, the Netherlands: Swets & Zeitlinger.

Bleichrodt, N., Resing, W. C. M., & Zaal, J. N. (1993). *Beoordeling schoolgedrag, SCHOBL-R: Handleiding en verantwoording* [School-Behavior Inventory, SCHOBL-R: Manual and justification]. Lisse, the Netherlands: Swets & Zeitlinger.

Davison, M. L., & Davenport, E. C. (2002). Identifying criterion-related patterns of predictor scores using multiple regression. *Psychological Methods, 7,* 468–483.

Douglas, J., & Cohen, A. (2001). Nonparametric function estimation for assessing parametric model fit. *Applied Psychological Measurement, 25,* 234–243.

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., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38,* 67–68.

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

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

Emons, W. H. M. (2003). Investigating the local fit of item-score vectors. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, & J. J. Meulman (Eds.). *New developments in psychometrics* (pp. 289–296). Tokyo: Springer.

Emons, W. H. M., Meijer, R. R., & Sijtsma, K. (2002). Comparing simulated and theoretical sampling distributions of the *U3* person-fit statistic. *Applied Psychological Measurement, 26,* 88–108.

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2004). Testing hypotheses about the person-response function in person-fit analysis. *Multivariate Behavioral Research, 39,* 1–35.

Fox, J. (1997). *Applied regression analysis, linear models, and related methods.* Thousand Oaks, CA: Sage.

Graham, J. R. (1993). *MMPI-2: Assessing personality and psychopathology* (2nd ed.). New York: University Press.

Grayson, D. A. (1988). Two group classification in latent trait theory: Scores with monotone likelihood ratio. *Psychometrika, 53,* 383–392.

Grossman, L. S., Haywood, T. W., & Wasyliw, O. E. (1988). The evaluation of truthfulness in alleged sex offenders' self reports: 16PF and MMPI validity scales. *Journal of Personality Assessment, 59,* 264–275.

Habing, B. (2001). Nonparametric regression and the parametric bootstrap for local dependence assessment. *Applied Psychological Measurement, 25,* 221–233.

Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika, 62,* 331–347.

Hoijtink, H., & Molenaar, I. W. (1997). A multidimensional item response model: Constrained latent class analysis using the Gibbs sampler and posterior predictive checks. *Psychometrika, 62,* 171–189.

Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *The Annals of Statistics, 14,* 1523–1543.

Junker, B. W. (1993). Conditional association, essential independence, and monotone unidimensional item response models. *The Annals of Statistics, 21,* 1359–1378.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16,* 277–298.

Karabatsos, G., & Sheu, C. F. (2004). Order-constrained Bayes inference for dichotomous models of unidimensional nonparametric IRT. *Applied Psychological Measurement, 28,* 110–125.

Klauer, K. C. (1991). An exact and optimal standardized person test for assessing consistency with the Rasch model. *Psychometrika, 56,* 213–228.

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.

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

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

Meijer, R. R. (2003). Diagnosing item score patterns on a test using item response theory-based person-fit statistics. *Psychological Methods, 8,* 72–87.

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

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

Mokken, R. J. (1971). *A theory and procedure of scale analysis.* Berlin, Germany: 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.

Molenaar, I. W., & Sijtsma, K. (2000). *MSP5 for Windows. User's manual.* Groningen, the Netherlands: ProGAMMA.

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

Pinsoneault, T. B. (2002). The clinical assessment of children and adolescents: A variable response inconsistency scale and a true response inconsistency scale for the Millon Adolescent Clinical Inventory. *Psychological Assessment, 14,* 320–330.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56,* 611–630.

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

Reise, S. P. (2000). Using multilevel logistic regression to evaluate person-fit in IRT models. *Multivariate Behavioral Research, 35,* 543–568.

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

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

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.

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

Sijtsma, K., & Junker, B. W. (1996). A survey of theory and methods of invariant item ordering. *British Journal of Mathematical and Statistical Psychology, 49,* 79–105.

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

Sijtsma, K., & Meijer, R. R. (2001). The person response function as a tool in person-fit research. *Psychometrika, 66,* 191–208.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory.* Thousand Oaks, CA: Sage.

Simonoff, J. S. (1996). *Smoothing methods in statistics.* New York: Springer-Verlag.

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.

Sprent, P. (1993). *Applied nonparametric statistical methods.* New York: Chapman & Hall.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52,* 589–617.

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

Tatsuoka, K. K., & Tatsuoka, M. M. (1983). Spotting erroneous rules by the Individual Consistency Index. *Journal of Educational Measurement, 20,* 221–230.

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

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

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

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

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

Zickar, M. J., & Drasgow, F. (1996). Detecting faking on a personality instrument using appropriateness measurement. *Applied Psychological Measurement, 20,* 71–88.

Received October 31, 2002
Revision received July 13, 2004
Accepted August 9, 2004 ■