The person response function as a tool in person-fit research

Sijtsma, K.; Meijer, R.R.

Published in:
Psychometrika

Publication date:
2001

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

PSYCHOMETRIKA—VOL. 66, NO. 2, 191–208
JUNE 2001

THE PERSON RESPONSE FUNCTION AS A TOOL IN PERSON-FIT RESEARCH

KLAAS SIJTSMA

TILBURG UNIVERSITY

ROB R. MEIJER

UNIVERSITY OF TWENTE

Item responses that do not fit an item response theory (IRT) model may cause the latent trait value to be inaccurately estimated. In the past two decades several statistics have been proposed that can be used to identify nonfitting item score patterns. These statistics all yield *scalar* values. Here, the use of the person response function (PRF) for identifying nonfitting item score patterns was investigated. The PRF is a *function* and can be used for diagnostic purposes. First, the PRF is defined in a class of IRT models that imply an invariant item ordering. Second, a person-fit method proposed by Trabin & Weiss (1983) is reformulated in a nonparametric IRT context assuming invariant item ordering, and statistical theory proposed by Rosenbaum (1987a) is adapted to test locally whether a PRF is nonincreasing. Third, a simulation study was conducted to compare the use of the PRF with the person-fit statistic ZU3. It is concluded that the PRF can be used as a diagnostic tool in person-fit research.

Key words: appropriateness measurement, invariant item ordering, nonparametric item response theory, person-fit method, person response function

Introduction

Person-fit research uses methods to identify respondents whose pattern of scores on the items from a test or a questionnaire is unusual, given the expectation based on a particular item response theory (IRT) model, or given the item score patterns produced by the majority of the respondents (e.g., Drasgow, Levine, & McLaughlin, 1987; Drasgow, Levine, & Zickar, 1996; Levine & Drasgow, 1982; Levine & Rubin, 1979; Meijer, 1996, 1998; Meijer & Sijtsma, 1995). A relatively rare approach to identifying aberrants is the use of the person response function (PRF), first discussed by Weiss (1973) and Lumsden (1978), and later discussed and applied by Trabin and Weiss (1983), Klauer and Rettig (1990), and Nering and Meijer (1998). The PRF, to be defined in greater detail later on, defines the probability of giving correct answers to dichotomous items as a function of an item difficulty scale. Trabin and Weiss chose the location parameter from the 3-parameter logistic model (3PLM; Lord, 1980); Klauer and Rettig discussed the PRF in the context of the Rasch (1960) model or 1-parameter logistic model (1PLM); and Lumsden discussed the PRF in a general IRT context. In general, it is assumed that the PRF is a *nonincreasing* function of item difficulty.

This study consists of three parts. First, we specify the desired properties of the PRF and provide its definition in a general, nonparametric IRT framework. To define the PRF as a nonincreasing function of item difficulty, we assume that the items have an invariant item ordering (IIO; Sijtsma & Junker, 1996); that is, we assume that item response functions (IRFs) do not intersect. Given an IIO, in a nonparametric IRT context a convenient choice for the item difficulty is 1 minus the proportion-correct on an item, which is well known from classical test theory.

The authors are grateful to Coen A. Bernaards for preparing the figures used in this article, and to Wilco H.M. Emons for checking the calculations.

Requests for reprints should be sent to Klaas Sijtsma, Department of Research Methodology, FSW, Tilburg University, PO Box 90153, 5000 LE Tilburg, The Netherlands. E-Mail: k.sijtsma@kub.nl

0033-3123/2001-2/1998-0667-A $00.75/0

Next, we discuss problems that arise when a nonparametric PRF is defined as a function of the latent IRT scale or when the PRF is defined in a logistic IRT context as a function of the latent item location parameter.

Second, we discuss the estimation of the PRF from empirical data. A method is presented for testing whether local deviations from the expectation that the PRF is a nonincreasing function are significant. The test is conservative, meaning that deviations have to be relatively large to be significant.

Third, a simulation study was conducted to compare the results of this local, conservative test with results of Van der Flier's (1982) *ZU3* person-fit statistic, which evaluates the entire item score pattern. Unlike ZU3 and other person-fit methods, which only allow for a *global* evaluation of the item score pattern, our PRF method provides the opportunity for monitoring the PRF and thus for detecting *local* deviations. This may render our method better suited for the diagnosis of aberrant behavior.

A Nonparametric Item Response Theory Framework

We consider tests consisting of J items. Each item is characterized by a binary item score variable, denoted X, and indexed $j = 1, \ldots, J$, so that $X_j = 0, 1$. A 0 score reflects an incorrect answer, and a 1 score reflects a correct answer. We assume that one latent trait denoted θ explains all dependencies between variables. The test thus is unidimensional. The conditional probability of obtaining a 1 score on item j is denoted $P_j(\theta) \equiv P(X_j = 1 \mid \theta)$. This conditional probability is the IRF. Further, we define a vector $\mathbf{X} = (X_1, \ldots, X_J)$ and a realization $\mathbf{x} = (x_1, \ldots, x_J)$. We assume that the item scores are locally independent,

$$P(\mathbf{X} = \mathbf{x} \mid \theta) = \prod_{j=1}^{J} P_j(\theta)^{x_j}[1 - P_j(\theta)]^{1-x_j}. \tag{1}$$

For any probability distribution of θ, denoted $F(\theta)$, θ can be integrated out of (1) which yields the joint marginal distribution of $\mathbf{X}$,

$$P(\mathbf{X} = \mathbf{x}) = \int_{\theta} \prod_{j=1}^{J} P_j(\theta)^{x_j}[1 - P_j(\theta)]^{1-x_j}\, dF(\theta). \tag{2}$$

Equation (2) does not restrict the distribution of $\mathbf{X}$ (see Suppes & Zanotti, 1981; also see Holland & Rosenbaum, 1986; and Junker, 1993). In order to have testable restrictions on the distribution of $\mathbf{X}$, specific choices for the $P_j(\theta)(j = 1, \ldots, J)$, for $F(\theta)$, or for both have to be made.

Nonparametric IRT models put order restrictions on the IRF, but refrain from a parametric definition of the IRF (Sijtsma, 1998). The first order restriction is that $P_j(\theta)$ is a nondecreasing function of θ (Ellis & van den Wollenberg, 1993; Junker, 1993; Mokken & Lewis, 1982; Stout, 1990). More specifically, for two arbitrarily chosen values of the latent trait, say $\theta_a < \theta_b$,

$$P_j(\theta_a) \leq P_j(\theta_b), \quad j = 1, \ldots, J. \tag{3}$$

The nonparametric IRT model based on (2) and (3) is the monotone homogeneity model (MHM; Mokken & Lewis, 1982; also see Ellis & van den Wollenberg, 1993; and Holland & Rosenbaum, 1986).

The practical importance of the MHM is that it implies (Hemker, Sijtsma, Molenaar, & Junker, 1997; based on a monotone likelihood ratio result by Grayson, 1988; and Huynh, 1994) that the latent trait θ is stochastically ordered by the unweighted sum of the scores on the J items

from the test, $X_+ = \Sigma X_j$. That is, for two values of X_+, say s and t, and any value of θ, say c,

$$P(\theta > c \mid X_+ = s) \leq P(\theta > c \mid X_+ = t), \quad \text{for all} \quad s < t. \tag{4}$$

Equation (4) also holds for unweighted sum scores based on subsets of the J items for which the MHM holds. Equation (4) implies that groups with higher X_+ scores have higher mean θs.

The second order restriction is that the IRFs are nonintersecting (Mokken & Lewis, 1982; Rosenbaum, 1987a, 1987b; Sijtsma & Junker, 1996). For a finite set of J items, all measuring the same unidimensional latent trait θ, we assume that the items can be ordered and numbered such that

$$P_1(\theta) \geq P_2(\theta) \geq \cdots \geq P_J(\theta), \quad \text{for all} \quad \theta. \tag{5}$$

Equation (5) is an order restriction on the J IRFs of the test. The IRFs do not intersect, which means that the ordering of the probabilities is the same, except for possible ties, for all values of θ. If items have an ordering as in Equation (5), they have a *latent scale* (Rosenbaum, 1987a) or an IIO (Sijtsma & Junker, 1996).

Unidimensionality, local independence, and nondecreasing and nonintersecting IRFs together define the double monotonicity model (DMM; Mokken & Lewis, 1982), which is a special case of the isotonic ordinal probabilistic model (ISOP; Scheiblechner, 1995). Equation (5) also holds for the 1PLM, the 1-parameter normal ogive model (Lord, 1952), and logistic and normal ogive models with all slope parameters fixed at the same value, and varying location and lower asymptote parameters which have opposite orderings (Sijtsma, 1998).

The Person Response Function

Properties of the PRF

Weiss (1973) and Lumsden (1978) suggested to consider the PRF to be the probability that person i ($i = 1, \ldots, n$) obtains a 1 score on an item (denoted by random variable $S_i = 1$) as a function of some item difficulty scale. With respect to the PRF, we assume that:

1. A continuous item location parameter δ' exists. Notation δ' is used to stress that δ' need not be equivalent with the location parameter δ from the logistic IRT models.
2. The probability that person i produces an item score of 1 is a function of δ'; this function is the PRF. By analogy with the IRF, $P(X_j = 1 \mid \theta)$, and fixed function values, $P(X_j = 1 \mid \theta_i)$, we define the PRF, $P(S_i = 1 \mid \delta')$, and fixed function values, $P(S_i = 1 \mid \delta'_j)$. The PRF gives the probability that person i provides correct answers to items measuring the same latent trait θ. For fixed item parameter δ', the PRF gives the probability that person i gives the correct answer to item j with location δ'_j.
3. The PRF is nonincreasing in δ'. This assumption seems to be reasonable given the interpretation of δ' as an item difficulty parameter and given the assumption of unidimensional measurement. The assumption of a nonincreasing PRF can be *unreasonable* if δ' does *not* order the items identically for each θ or if the items are *multidimensional*. The first possibility is discussed later on. As regards the second assumption, assume a fixed θ_i and $\delta'_j < \delta'_k$, and an ordering of probabilities such that $P_i(\delta'_j) < P_i(\delta'_k)$; this means that the PRF is not a monotonely nonincreasing function. An explanation for this result is that the items measure different traits (see Trabin & Weiss, 1983, pp. 91–92).

Summary. The PRF describes the probability that person i with θ_i gives positive answers (scored $S_i = 1$) to items measuring θ. The PRF is monotonically nonincreasing in an item location parameter δ' (to be defined shortly) and is denoted $P_i(\delta') \equiv P(S_i = 1 \mid \delta')$.

A Nonparametric PRF Definition Based on an Observable Scale

The nonparametric IRT framework provides an opportunity for a useful PRF definition. Given that an IIO (Equation (5)) holds, a PRF can be defined as a function of the observable item proportions

$$\pi_j = \int_\theta P_j(\theta)\, dF(\theta). \tag{6}$$

The corresponding sample fraction is $\hat{\pi}_j = n_j/n$; n_j is the sample frequency with item j correct and n is the sample size. From Equations (5) and (6) it readily follows that

$$\pi_1 \geq \pi_2 \geq \cdots \geq \pi_J. \tag{7}$$

Moreover, Sijtsma and Molenaar (1987, Theorem 1) showed that under the DMM, of which IIO (Equation (5)) is an assumption, items have equal proportions $\pi_j = \pi$ if and only if the IRFs coincide:

$$P_1(\theta) = \cdots = P_J(\theta) \Longleftrightarrow \pi_1 = \cdots = \pi_J. \tag{8}$$

Corollary 1. Assume that J items have an IIO. Also, assume that whenever two IRFs do not coincide in a particular interval of the θ scale, the probability density $f(\theta) > 0$ for θs from this interval. Then, if none of the J IRFs coincides completely with any of the other $J-1$ IRFs, each item has a unique π_j value, and the strict ordering of πs,

$$\pi_1 > \pi_2 > \cdots > \pi_J,$$

uniquely describes the item ordering by $P_j(\theta)$, except for θs for which this ordering (i.e., by $P_j(\theta)$) contains ties (see (5)).

Proof. Suppose that two items j and k ($j < k$) have an IIO, and that their IRFs coincide with the exception of T arbitrarily narrow θ intervals, enumerated $[\theta]_1, [\theta]_2, \ldots, [\theta]_T$. Note that

$$\pi_j - \pi_k = \int_\theta [P_j(\theta) - P_k(\theta)]\, dF(\theta). \tag{9}$$

If $T = 0$, then for all θs the difference between brackets is 0, which means that $\pi_j = \pi_k$ (Equation (8)). If $T \geq 1$, then by notational convention in each of the intervals $[\theta]_1, [\theta]_2, \ldots, [\theta]_T$ the difference between brackets in (9) is positive; consequently, $\pi_j > \pi_k$. The generalization to any J follows trivially. □

We use π_j as a candidate for the item difficulty parameter generically denoted δ'. Thus, we may define the probability that person i (with θ_i) has a 1 score on item j as $P_i(\pi_j)$ [instead of $P_i(\delta'_j)$]. For J items having an IIO, we then have

$$P_i(\pi_1) \geq P_i(\pi_2) \geq \cdots \geq P_i(\pi_J). \tag{10}$$

Because increasing $1 - \pi_j$ means higher difficulty, we define $\delta' \equiv 1 - \pi$.

Definition. For the continuous scale of $1 - \pi$ with domain [0, 1], we define the PRF to be $P(S_i = 1 \mid 1 - \pi)$, which is a *nonincreasing* function provided that an IIO holds.

Figure 1 shows three theoretical PRFs satisfying the Definition.

It may be noted that other IRT models, which assume unidimensionality, local independence, and nondecreasing IRFs, and which imply an IIO may give rise to useful PRF definitions. For example, in the 1PLM the ordering of the location parameters $\delta_1 < \delta_2 < \cdots < \delta_J$ reflects

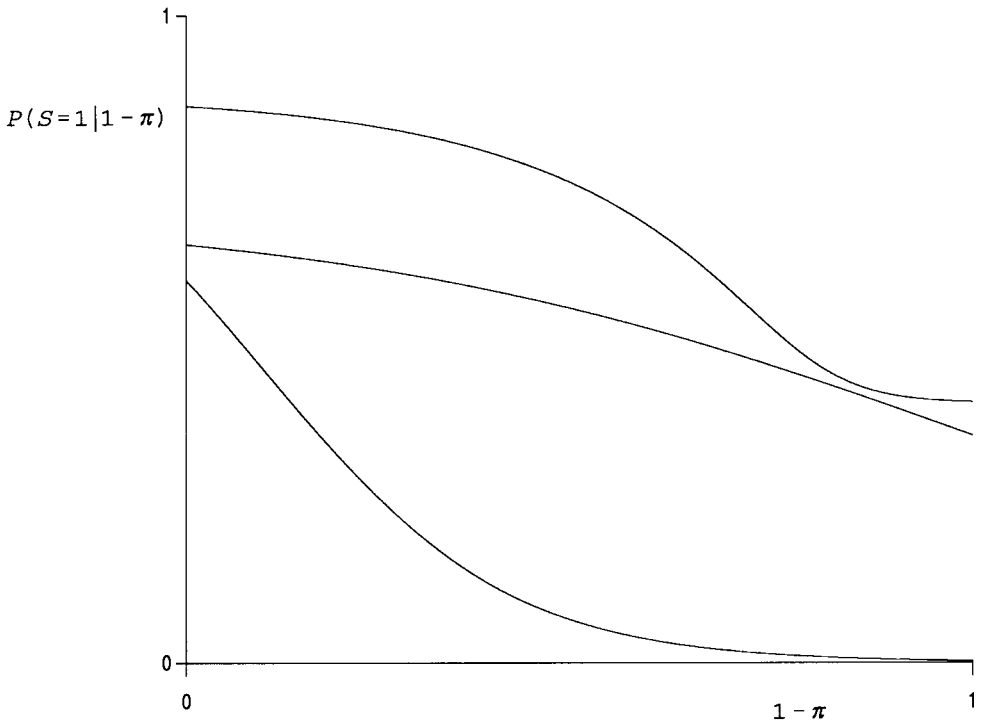

FIGURE 1.
Three nonincreasing person response functions $P(S = 1 \mid 1 - \pi)$ defined under models implying an invariant item ordering.

the reverse item ordering by $P_1(\theta) > P_2(\theta) > \cdots > P_J(\theta)$. This result is in the spirit of Corollary 1 and implies a nonincreasing PRF as a function of $\delta' \equiv \delta$. A similar result holds for the 1-parameter normal-ogive model (Lord, 1952).

IRT models not implying an IIO and other definitions of the generic item difficulty δ' lead to PRFs which are either nonmonotonous or ill-defined (because for some items the parameter δ' is undefined). Because this topic has not been explored in great depth in the literature, the next section gives some results illustrating these points.

Problematic PRF Definitions

In our discussion of some of the problems arising with other PRF definitions, we need the 4-parameter logistic model (4PLM; Hambleton & Swaminathan, 1985, pp. 48–49) and its special cases, the 1-, 2-, and 3PLM. The 4PLM provides a general parametric definition of the IRF. The 4PLM has logistic IRFs, which vary in location δ, slope α, lower asymptote γ for $\theta \to -\infty$, and upper asymptote λ for $\theta \to \infty$ ($\gamma_j \leq \lambda_j \leq 1$). By including the upper asymptote parameter, the possibility is modeled that for some items the probability of a correct answer is lower than 1 even for the highest-ability respondents in a particular population. The 4PLM is defined as

$$P_j(\theta) = \gamma_j + \frac{(\lambda_j - \gamma_j)\exp[\alpha_j(\theta - \delta_j)]}{1 + \exp[\alpha_j(\theta - \delta_j)]}. \tag{11}$$

The 3PLM is obtained by fixing $\lambda_j = 1$, for all j; the 2PLM is obtained by additionally fixing $\gamma_j = 0$, for all j; and the 1PLM is obtained by additionally fixing $\alpha_j = 1$, for all j.

Problems When a Nonparametric PRF Definition Is Based on the Latent θ Scale

Suppose we adopt the definition of $\delta' \equiv \delta$ from the logistic IRT models (Equation (11)) in a nonparametric context. In logistic models, δ_j is the value of θ such that the probability $P_j(\delta)$ lies exactly halfway the lowest and the highest possible probabilities for item j. For the 1PLM and the 2PLM, $P_j(\delta_j) = \frac{1}{2}$; for the 3PLM, $P_j(\delta_j) = \frac{1}{2}(\gamma_j + 1)$; and for the 4PLM, $P_j(\delta_j) = \frac{1}{2}(\gamma_j + \lambda_j)$. Sijtsma and Junker (1996) showed that if this definition of δ is adopted in the nonparametric framework, defined by (2), (3), and (5), in which any form of a nondecreasing function is allowed within the constraints of nondecreasingness and nonintersection, the ordering of the items by δ may, for example, suggest that item 1 is more difficult than item 2 ($\delta_1 > \delta_2$) although $P_1(\theta) > P_2(\theta)$, for all θ, leads to the opposite conclusion (Figure 2).

Alternatively, we may define a location parameter $\delta' \equiv \delta^*$ as the value of θ for which $P_j(\delta^*) = d$, where d is a constant. For example, for $d = 0.5$ Figure 3 shows that the item ordering by δ^* and the item ordering by $P_j(\theta)$, for all θ, are opposite [$\delta_2^* < \delta_3^*$; $P_2(\theta) > P_3(\theta)$], thus suggesting the same difficulty ordering (also see Figure 2). However, the choice of d is not without problems. If d is low, say $d = d_L$, δ^* is undefined for IRFs with a minimum value higher than d_L; and if d is high, say $d = d_H$, δ^* is undefined for IRFs with a maximum value lower than d_H. Moreover, if the highest minimum (denoted $P_{\text{min(sup)}}$) of all J IRFs is higher than the lowest maximum (denoted $P_{\text{max(inf)}}$), no choice of d exists such that δ^* is defined for each IRF from the test. For $J = 4$, Figure 3 shows a situation where $P_{\text{min(sup)}} > P_{\text{max(inf)}}$, and it is easily seen that no choice of d exists that yields values for all δ_j^*, $j = 1, \ldots, J$. For example, if $d = 0.5$, only δ_2^* and δ_3^* are defined.

Obviously, if $P_{\text{min(sup)}} \leq P_{\text{max(inf)}}$, values of d such that

$$P_{\text{min(sup)}} \leq d \leq P_{\text{max(inf)}} \tag{12}$$

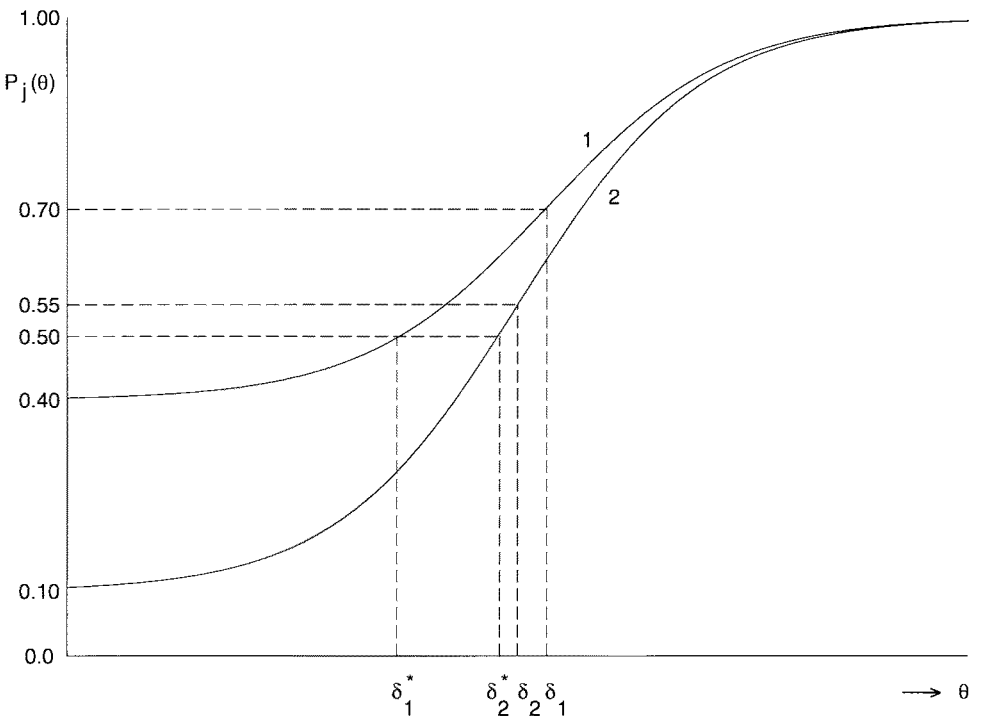

FIGURE 2.
Two item response functions with $P_1(\theta) > P_2(\theta)$ for all θs, and $\delta_1 > \delta_2$ and $\delta_1^* < \delta_2^*$.

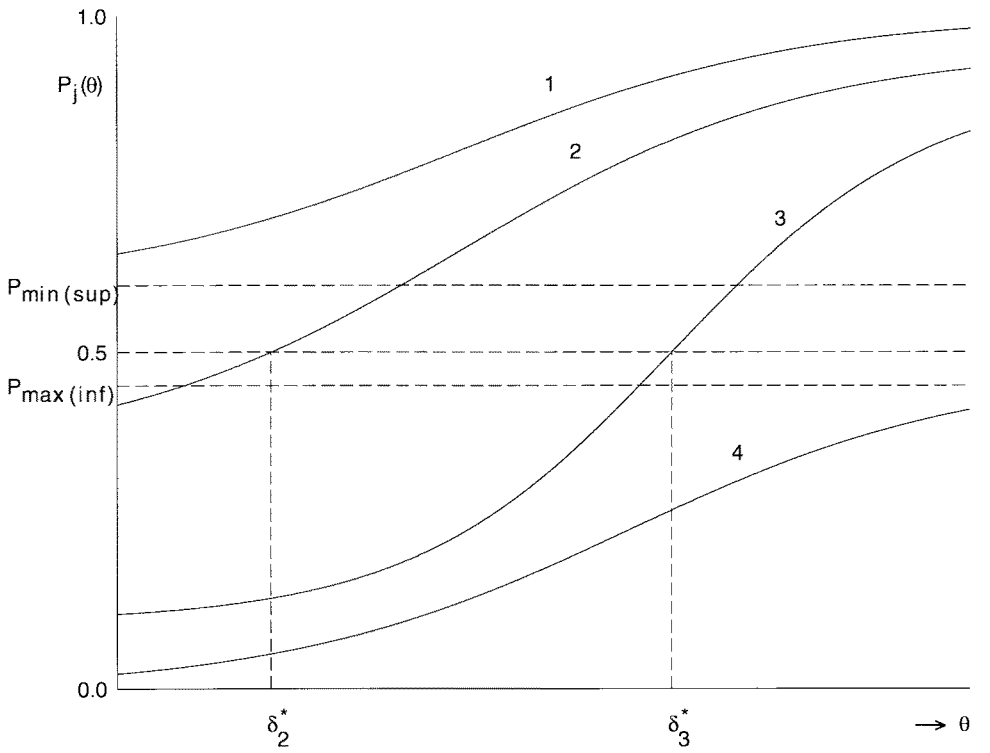

FIGURE 3.
Four item response functions with $P_{\text{min(sup)}} > P_{\text{max(inf)}}$ so that no choice of d exists such that locations δ^* are defined for all items.

always yield parameters δ_j^* for each $j = 1, \ldots, J$; other choices of d deprive one or more items of a location δ^*. Given (12), any choice of d leads to an ordering of items by δ^* that is opposite, except for possible ties, to the item ordering by $P_j(\theta)$ in (5). This means that both orderings yield the same difficulty ordering, except for possible ties. Because (12) obviously restricts the definition of a PRF as a function of a latent item difficulty, and because we prefer a definition which is generally applicable, we refrain from further exploring the use of an item parameter defined on the latent θ scale.

Problems With the PRF Definition in Some Parametric Models

In the 2PLM, the 3PLM, and the 4PLM, the definition of a PRF as a function of the location parameter δ is problematic because IRFs from these models intersect. For example, unless $\alpha_j = \alpha_k$ in the 2PLM two IRFs intersect at

$$\theta_{jk} = \frac{\alpha_j \delta_j - \alpha_k \delta_k}{\alpha_j - \alpha_k}.$$

Given that $\alpha_j \neq \alpha_k$, the ordering of δ_j and δ_k matches two orderings of probabilities: if $\alpha_j < \alpha_k$, then for $\theta < \theta_{jk}$ we have $P_j(\theta) > P_k(\theta)$, and for $\theta > \theta_{jk}$ we have $P_j(\theta) < P_k(\theta)$; and if $\alpha_j > \alpha_k$ the orderings are opposite. In general, if all J items have different slope parameters, then their IRFs have $\frac{1}{2}J(J-1)$ intersection points and define $\frac{1}{2}J(J-1)+1$ different orderings of conditional probabilities. Clearly, the ordering of J items according to δ represents *only one* of the orderings according to $P(\theta)$. In general, under these models a PRF (defined for fixed θ) is not a nonincreasing function of δ.

The PRF is only a nonincreasing function of δ for nonintersecting subsets of IRFs. For example, two 3PLM IRFs with slope parameters fixed at 1, lower asymptotes $\gamma_j > \gamma_k$ and locations $\delta_j < \delta_k$ do not intersect. This is shown by first noting that they would intersect if roots exist for

$$(\gamma_j - \gamma_k) + (\gamma_j - 1)\exp(\theta - \delta_k) + (1 - \gamma_k)\exp(\theta - \delta_j) = 0. \tag{13}$$

For the parameter setup chosen here, we have

$$\gamma_j > \gamma_k \Longrightarrow 1 - \gamma_k > |\gamma_j - 1|;$$

and

$$\delta_j < \delta_k \Longrightarrow \exp(\theta - \delta_k) < \exp(\theta - \delta_j).$$

It follows that the left-hand side of (13) is positive. Second, if $\gamma_j = \gamma_k$ and $\delta_j = \delta_k$, then $P_j(\theta) = P_k(\theta)$, for all θ. As a result, we have shown that the IRFs do not intersect. In general, J 3PLM IRFs do not intersect if (1) $\alpha_1 = \alpha_2 = \cdots = \alpha_J$ and (2) $\gamma_1 > \gamma_2 > \cdots > \gamma_J$ and $\delta_1 < \delta_2 < \cdots < \delta_J$. Thus, for each value of θ, say θ_i, $P_i(\delta)$ is nonincreasing. Similarly, in the 4PLM subsets of nonintersecting IRFs can be defined. Because we are interested in a generally useful PRF definition, the identification of nonintersecting sets of IRFs under the 3PLM and the 4PLM is not further pursued.

A Nonparametric Person-Fit Approach Based on IIO That Uses the PRF

Estimation of the PRF

Trabin and Weiss (1983; also see Nering & Meijer, 1998) defined a PRF in the context of the 3PLM, using the δ parameter. They discussed how the PRF may provide information about an individual's carelessness (too low success probability on easy items), guessing tendency (too high success probability on difficult items), and accuracy (steepest slope downward; cf. α for the logistic IRF). We adapt the PRF definition proposed by Trabin and Weiss (1983) to the nonparametric context.

Let J items be ordered and numbered, such that

$$1 - \pi_1 \leq 1 - \pi_2 \leq \cdots \leq 1 - \pi_J. \tag{14}$$

J is chosen such that G ordered classes A_g $(g = 1, \ldots, G)$ can be formed, each containing m items; thus, $A_1 = \{1, \ldots, m\}$, $A_2 = \{m + 1, \ldots, 2m\}, \ldots, A_G = \{J - m + 1, \ldots, J\}$. Within each class, for θ_i the expected proportion of correct answers is obtained through

$$\pi_{ig} = m^{-1} \sum_{j \in A_g} P_j(\theta_i), \quad \text{for} \quad g = 1, \ldots, G. \tag{15}$$

By assuming IIO, for any pair of IRFs from adjacent subsets g and $g + 1$ we have

$$P_j(\theta_i) \geq P_{j'}(\theta_i), \quad \text{for} \quad j \in A_g \quad \text{and} \quad j' \in A_{g+1}, \tag{16}$$

and, thus, $\pi_{ig} \geq \pi_{i,g+1}$. Generalizing this result to G ordered item subsets we have

$$\pi_{i1} \geq \pi_{i2} \geq \cdots \geq \pi_{iG}. \tag{17}$$

This is a discrete (J finite) approximation to the PRF. Equation (17) is estimated as follows. Since $E(X_j \mid \theta) = P_j(\theta)$, the score on item j is taken as a binary estimate of the success probability.

For person i who has item scores $X_{ij} = x_{ij}$, the sample fraction $\hat{\pi}_{ig}$ is

$$\hat{\pi}_{ig} = m^{-1} \sum_{j \in A_g} X_{ij}, \quad \text{for} \quad g = 1, \ldots, G. \tag{18}$$

Given (17), a sample ordering

$$\hat{\pi}_{i1} \geq \hat{\pi}_{i2} \geq \cdots \geq \hat{\pi}_{iG} \tag{19}$$

definitely supports a nonincreasing PRF, but significant deviations from the expected ordering may give evidence of person-misfit.

Testing for Nonincreasingness of the PRF

Let 80 items be divided into $G = 8$ subsets of 10 items each, such that A_1 contains the 10 easiest items, A_2 contains the next ten easiest items, and so on. As an example, consider A_1 and A_2, and assume that person i has scores {(1010000100), (1111011011)}. Let X_+ denote the total score on all items from A_1 and A_2; X_{+e} the total score on the relatively easy subset (i.e., A_1); J the number of items in A_1 and A_2; and J_e the number of items in the easy subset. A useful question in person-fit analysis is whether $X_{+e} = 3$ is exceptionally low, given that $J = 20$, $J_e = 10$, and $X_+ = 11$.

To answer this question, we derive a corollary based on a theorem by Rosenbaum (1987a). More specifically, we derive a conservative bound for the local significance test that for item subsets A_g and $A_{g'} (g < g')$ the total score X_{+e} is exceptionally low, given J, J_e, and X_+. The theory developed here assumes scoring functions which are *decreasing in transposition* (DT; see Rosenbaum, 1987a, for the definition). Scoring functions such as X_+ (and X_{+e}) are DT functions.

Our corollary is given for the subdivision of a test into three parts. Let J be the number of items in the entire test and let the J item scores be collected in $\mathbf{X}$. We define $\mathbf{X} = (\mathbf{Y}, \mathbf{Z}) = (\mathbf{Y}_e, \mathbf{Y}_d, \mathbf{Z})$. $\mathbf{Y}_e$ contains the J_e easiest items from $\mathbf{Y}$ and $\mathbf{Y}_d$ contains the J_d most difficult items; $J_e + J_d = J_Y$. We assume that the items in $\mathbf{Y}$ have an IIO (Equation (5)). Let $\mathrm{f}(\mathbf{Y}_e) = X_{+e}$ be the unweighted number-correct score on the J_e items from $\mathbf{Y}_e$; $\mathrm{f}(\mathbf{Y}_e)$ is DT. The items in $\mathbf{Z}$ can be used for selecting subgroups of respondents, for example, all respondents having fewer than x_{+z} items correct.

If for each θ all $\binom{J_Y}{x_+}$ possible item score patterns with x_+ 1 scores would have equal probability, then X_{+e} would follow the hypergeometric distribution given J_Y, J_e, and X_+, and $P(X_{+e} \leq x_{+e} \mid J_Y, J_e, X_+)$ would be the probability of interest. Because we assumed that all item score patterns have equal probability, $\mathbf{Y}$ follows the *exchangeable* distribution (Lindgren, 1993).

Rosenbaum (1987a) compares the expectation of the DT function $\mathrm{f}(\mathbf{Y})$ *given IIO* with the expectation of $\mathrm{f}(\mathbf{Y})$ *given the exchangeable distribution*. We define the indicator function $I[\mathrm{f}(\mathbf{Y})] = 1$ if $\mathrm{f}(\mathbf{Y}) \geq c^* (1 \leq c^* \leq J_y - 1)$; and 0 otherwise. $I[\mathrm{f}(\mathbf{Y})]$ is DT whenever $\mathrm{f}(\mathbf{Y})$ is DT (see Rosenbaum, 1987a). Our corollary says that the probability that $I[\mathrm{f}(\mathbf{Y})] = 1$ is higher under an IIO than under the exchangeable distribution.

Corollary 2. If $\mathbf{X} = (\mathbf{Y}, \mathbf{Z})$ has a latent variable representation (Equation (2)), and if the items in $\mathbf{Y}$ have an IIO, then for the indicator function $I[\mathrm{f}(\mathbf{Y})]$, and for any arbitrary function $\mathrm{h}(.)$,

$$P\{I[\mathrm{f}(\mathbf{Y})] = 1 \mid X_+ = x_+, \mathrm{h}(\mathbf{Z})\} \geq P\{I[\mathrm{f}(\mathbf{Q})] = 1\}; \tag{20}$$

$\mathbf{Q}$ has J_Y elements (0, 1); and $\mathbf{Q}$ has the exchangeable distribution; thus,

$$P(\mathbf{Q} = \mathbf{q}) = \binom{J_Y}{x_+}^{-1}.$$

If $\mathbf{Y} = (\mathbf{Y}_e, \mathbf{Y}_d)$, then for $f(\mathbf{Y}) = X_{+e}$ (20) implies that given IIO the probability of obtaining at least $X_{+e} = c^*$ 1s is at least as high as under the exchangeable distribution. Reversely, it follows that under an IIO the probability of obtaining *at most* $X_{+e} = c^*$ 1's cannot be higher than under the exchangeable distribution. Thus, the exchangeable distribution provides a conservative bound on the probability $P(X_{+e} \leq x_{+e})$. This bound is obtained from $P(X_{+e} \leq x_{+e} \mid J_Y, J_e, X_+)$, which is found from the cumulative hypergeometric distribution.

How can Corollary 2 be used for testing hypotheses about the PRF? Assume that all J item score variables are collected in $\mathbf{Y}$ (i.e., $\mathbf{Z}$ is empty). Let subsets A_g of m increasingly more difficult items be collected in exhaustive and mutually exclusive vectors $\mathbf{Y}_g$, such that $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_G)$. Consider newly defined vectors $\mathbf{Y}_{(g)}$, each of which contains two adjacent subsets A_g and A_{g+1}: $\mathbf{Y}_{(1)} = (\mathbf{Y}_1, \mathbf{Y}_2), \mathbf{Y}_{(2)} = (\mathbf{Y}_2, \mathbf{Y}_3), \ldots, \mathbf{Y}_{(G-1)} = (\mathbf{Y}_{G-1}, \mathbf{Y}_G)$. Corollary 2 separately applies to each pair in the $\mathbf{Y}_{(g)}$'s. Thus, for each pair a conservative bound based on the hypergeometric distribution can be calculated for the probability that a person has at most $X_{+e} = x_{+e}$ 1s on the easiest item subset. If for a particular pair this probability is lower than, say, 0.05, then the conclusion is that the total score (or a lower score) on the first subset in this pair is unlikely, given that all items in the first subset are easier than the items in the second subset.

The cumulative hypergeometric probability $\mathcal{P}$ has to be calculated bearing in mind that if $X_+ > J_d$ the minimum possible value of X_{+e} is $X_+ - J_d$. Thus,

$$\mathcal{P} \equiv P(X_{+e} \leq x_{+e} \mid J, J_e, X_+) = \sum_{w=\max(0, X_+ - J_d)}^{x_{+e}} P(X_{+e} = w \mid J, J_e, X_+). \qquad (21)$$

For the example given previously, based on $X_{+e} = 1, 2, 3$, $\mathcal{P} = 0.035$. By Corollary 2 the real probability is lower, implying that at a 5% significance level we have sufficient evidence that the PRF increases at the easy part of the item difficulty scale. At a 1% level, however, the evidence is insufficient because the real probability might be higher than 0.01.

Simulation Study

The main purpose of this simulation study was to explore the usefulness of the cumulative hypergeometric upper bound $\mathcal{P}$ (Equation (21)) for detecting aberrant PRFs. Upper bound $\mathcal{P}$ was used for determining the detection rate using the PRF approach, and was compared with the detection rate of the nonparametric ZU3 person-fit statistic (Van der Flier, 1982). ZU3 is a standardized person-fit statistic which is asymptotically standard normally distributed.

We used the 4PLM for simulating data in agreement with IIO by choosing appropriate restrictions on the item parameters α, δ, γ, and λ (Equation (11)). The 4 PLM provides more flexibility for simulating IIO than the other logistic IRT models. In the nonparametric context IRFs typically can be estimated using nonparametric regression methods (e.g., Ramsay, 1991, 1995), but for generating data these estimates do not easily provide us with the response probabilities needed for simulating 0's and 1's. Thus, in the nonparametric context the 4PLM is a model with enough flexibility which also provides response probabilities for each simulee on each item.

Method

Independent Variables

Earlier person-fit research (e.g., Klauer, 1991; Meijer, Molenaar, & Sijtsma, 1994; Molenaar & Hoijtink, 1990) showed that the detection rate of a person-fit statistic is a function of (a) the test length, (b) the model that describes the test data, (c) the way the item scores deviate from the model, (d) the nominal significance level, and (e) the way a test is subdivided into two or more subtests. A design including these factors was used to evaluate the usefulness of the cumulative hypergeometric $\mathcal{P}$ (Equation (21)) for detecting aberrant PRFs.

Test length. Tests of 40 and 80 items were used. These numbers correspond to moderate and long tests, respectively. Meijer et al. (1994) concluded that person-fit research is not recommendable for shorter tests because of the low detection rate. Longer unidimensional tests or subtests from test batteries seem to be relatively rare in practice.

IRT model. The models used to describe the data were the 1PLM and the more complex 4PLM with λ and α both fixed at constant values, and oppositely ordered γ and δ parameters (thus, $\gamma_j > \gamma_k$ and $\delta_j < \delta_k$, for $j < k$). This configuration of parameters implies an IIO. To generalize the findings beyond IIO, additional datasets were simulated under IRT models that do not imply an IIO (see below).

Model violations. We simulated carelessness (see, e.g., Trabin & Weiss, 1983, p. 91). Carelessness may result in answering fewer items correct in the beginning of the test than expected on the basis of an examinee's ability level. Other forms of aberrant behavior, for example, cheating and guessing, may manifest themselves in more unpredictable places in the test. In the present simulation study, it was convenient to know where aberrance could be expected.

Nominal significance level. The nominal significance levels were 0.10 and 0.05.

Division into subtests. The 40-item test was divided into two, four, and eight subtests, each containing 20, 10, and 5 items, respectively. The 80-item test was divided into two, four, and eight subtests, each containing 40, 20, and 10 items, respectively. The division of a test into two halves yields only two points for estimating the PRF. Obviously, this situation has to be considered as an extreme case.

Simulation Procedure and Dependent Variables

Itemsets having an IIO. Item scores were generated as follows. The 1PLM and the 4PLM were used for generating item scores. For the 1PLM, datasets of 3000 model-fitting item score vectors were generated separately for $\theta = -2, -1, 0, 1, 2$; and for each θ using equidistant δ from U(−2, 2). This was done both for the 40-item test and the 80-item test. For 4PLM items having an IIO, two different configurations of α and γ parameters were used. For the 40-item test:

1. Datasets (N = 3000) were generated separately for $\theta = -2, -1, 0, 1, 2$. For each θ, $\alpha = 1$ and $\lambda = 0.8$ for all items; furthermore, $\delta_1 = -2.0, \delta_2 = -1.9, \delta_3 = -1.8, \ldots, \delta_{40} = 2.0$, with $\delta = 0$ excluded to obtain a symmetrical distribution around 0; and $\gamma_1 = 0.40, \gamma_2 = 0.39, \gamma_3 = 0.38, \ldots, \gamma_{39} = 0.02, \gamma_{40} = 0.01$;
2. Datasets (N = 3000) were generated separately for each $\theta = -2, -1, 0, 1, 2$. For each θ, $\alpha = 1$ for all items; and three subsets of items were distinguished, with $\lambda = 0.9$ for all items in subset 1, $\lambda = 0.8$ for all items in subset 2, and $\lambda = 0.7$ for all items in subset 3. Subset 1 consisted of 20 items with $\delta_1 = -2.0, \delta_2 = -1.9, \ldots, \delta_{20} = -0.1$; and $\gamma_1 = 0.40, \gamma_2 = 0.39, \ldots, \gamma_{20} = 0.21$. Subset 2 consisted of 10 items with $\delta_{21} = 0.1, \delta_{22} = 0.2, \ldots, \delta_{30} = 1.0$; and $\gamma_{21} = 0.20, \gamma_{22} = 0.19, \ldots, \gamma_{30} = 0.11$. Subset 3 consisted of 10 items with $\delta_{31} = 1.1, \delta_{32} = 1.2, \ldots, \delta_{40} = 2.0$; and $\gamma_{31} = 0.10, \gamma_{32} = 0.09, \ldots, \gamma_{40} = 0.01$.

Compared to the 40-item test, within the 80-item test for each item subset the number of items was doubled, and values for the δ's and γ's were chosen by keeping the range the same as for the 40-item test and halving the distance between adjacent parameter values.

After data had been generated using the 1PLM or the 4PLM, misfitting item score patterns were generated for all θ's. To simulate carelessness, the probability of correctly answering an item was set to 0.25 on the 5 easiest items of the 40-item test and the 10 easiest items of the 80-item test.

The complete itemset was then subdivided into two, four, or eight subtests of increasing difficulty, and for each simulee the cumulative hypergeometric $\mathcal{P}$ was determined for pairs of adjacent subtests. For example, for the 40-item test and four increasingly difficult subtests

$\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3, \mathbf{Y}_4)$, the probabilities of exceedence were determined for $\mathbf{Y}_{(1)} = (\mathbf{Y}_1, \mathbf{Y}_2)$, $\mathbf{Y}_{(2)} = (\mathbf{Y}_2, \mathbf{Y}_3)$, and $\mathbf{Y}_{(3)} = (\mathbf{Y}_3, \mathbf{Y}_4)$. For the subtests under consideration, the detection rate was the percentage of simulees with $\mathcal{P}$ values below nominal significance levels of 0.05 and 0.10, respectively. Since carelessness only took place on the easiest items in $\mathbf{Y}_1$, aberrant total scores were expected only when comparing $\mathbf{Y}_1$ and $\mathbf{Y}_2$. Furthermore, for each simulee ZU3 was determined using all 40 or 80 items in the test, and the detection rate was the percentage of simulees with a ZU3 value higher than 1.65 (one-tailed 0.05 error rate) and 1.29 (one-tailed 0.10 error rate).

Itemsets not having an IIO. The robustness of the cumulative hypergeometric $\mathcal{P}$ was investigated under mild violations of IIO for the 40- and 80-item test; $\theta = -2, 0$, and 2; the 2PLM with α drawn at random from U(.8, 1.2) and the 4PLM with $\lambda = 0.8$ for all items, γ's equidistant between 0.40 and 0.01 (40 γ's for the 40-item test and 80 γ's for the 80-item test), and α drawn at random from U(.8, 1.2). This choice of α's violated IIO.

Results

Results for cumulative hypergeometric $\mathcal{P}$, for 40 items. Table 1 gives the proportions of simulees with $\mathcal{P}$ values (Equation (21)) lower than 0.05 and 0.10. These proportions are given for different θ values, for: (a) the 1PLM and two configurations of the item parameters of the 4PLM; (b) test length of 40 items; and (c) carelessness on the five easiest items.

For the 1PLM (Table 1, left panel), the subdivision of the test into two subtests of 20 items each, for $\mathbf{Y}_1 - \mathbf{Y}_2$ resulted in detection rates equal to 0 or near 0 ($\theta = 2$: .004 and .005). The division of the test into four subtests of 10 items each, for nominal significance level of 0.05 resulted in detection rates ranging from .003 ($\theta = -2$) to .347 ($\theta = 2$). For nominal significance level of 0.10, detection rates were somewhat higher. For other subtest groupings $\mathbf{Y}_2 - \mathbf{Y}_3$ and $\mathbf{Y}_3 - \mathbf{Y}_4$ (not tabulated) detection rates were .001, which was expected based on how carelessness was implemented. The subdivision of the test into eight subtests of 5 items each, for $\mathbf{Y}_1 - \mathbf{Y}_2$ and nominal significance level of 0.05 resulted in detection rates between .018 and .584. For the other adjacent groupings almost no significant differences were found (not tabulated). The subdivision of the test into eight subtests meant that all five item scores in $\mathbf{Y}_1$ were aberrant; this yielded higher detection rates than for the other subdivisions. For each subdivision, detection rates increased in θ, because the probability of correct answers increased in θ. Thus, deviations from the expected number-correct in $\mathbf{Y}_1$ tended to be larger for higher θ's.

Compared with the 1PLM, the 4PLM with a fixed upper asymptote ($\lambda = 0.8$; Table 1, middle panel) resulted in lower detection rates for subdivisions into four and eight subtests for $\theta = 0, 1, 2$. The difference with the 1PLM was smaller for the subdivision into four subtests. The larger difference for the subdivision into eight subtests can be explained by the higher mean proportion-correct (Equation (6)) of the items in $\mathbf{Y}_2$ when using the 1PLM. For example, for the items in $\mathbf{Y}_2$, under the 1PLM for $\theta = 1$ these proportions ranged from 0.93 to 0.84 (with increasing δ), and for the 4PLM the corresponding proportions ranged from 0.77 to 0.75. As a result, in general the person number-correct score X_{+d} in $\mathbf{Y}_2$ was higher under the 1PLM than under the 4PLM, and aberrant person number-correct scores X_{+e} in $\mathbf{Y}_1$ were more likely under the 1PLM than under the 4PLM. For $\theta = -2, -1$, the detection rates under the 4PLM were somewhat higher than under the 1PLM due to the somewhat higher item proportion-correct score in $\mathbf{Y}_2$ under the 4PLM. For $\theta = 0, 1, 2$, the detection rates for the 4PLM with varying upper asymptotes (Table 1, right panel) for most cells often were in between the results for the 1PLM and the 4PLM with a fixed upper asymptote for all items. For $\theta = -2, -1$, the results were comparable with the 4PLM with a fixed upper asymptote.

Results for ZU3, for 40 items. For all three IRT models and for all five θ's and at both significance levels, for ZU3 detection rates were higher than for $\mathcal{P}$. For example, for the 1PLM and $\theta = 0$ the detection rate was 0.723 when ZU3 $>$ 1.29 was used, and 0.626 when ZU3 $>$ 1.65 was used; corresponding highest detection rates using $\mathcal{P}$ were 0.451 and 0.302, respectively.

TABLE 1.
Detection rate using the cumulative hypergeometric $\mathcal{P}$ and ZU3 for careless simulees on the 5 easiest items of the 40-item test. Detection rates are given only for $\mathbf{Y}_1 - \mathbf{Y}_2$ subtests

1PLM						4PLM[1]						4PLM[2]					
Length subtest/#subtests						Length subtest/#subtests						Length subtest/#subtests					
θ	sig. lev.	20/2	10/4	5/8	ZU3	θ	sig. lev.	20/2	10/4	5/8	ZU3	θ	sig. lev.	20/2	10/4	5/8	ZU3
−2	.05	0	.003	.018	.133	−2	.05	0	.014	.057	.162	−2	.05	0	.013	.052	.212
	.10	0	.006	.062	.204		.10	0	.036	.146	.280		.10	0	.049	.139	.340
−1	.05	0	.012	.096	.284	−1	.05	0	.025	.115	.230	−1	.05	0	.024	.149	.267
	.10	0	.040	.174	.420		.10	.001	.077	.214	.362		.10	0	.079	.273	.423
0	.05	0	.048	.302	.626	0	.05	0	.049	.193	.289	0	.05	0	.056	.260	.402
	.10	0	.112	.451	.723		.10	.001	.124	.283	.429		.10	0	.136	.391	.562
1	.05	0	.134	.450	.831	1	.05	.003	.103	.255	.307	1	.05	0	.123	.371	.487
	.10	0	.256	.652	.912		.10	.004	.204	.402	.459		.10	0	.192	.513	.630
2	.05	.004	.347	.584	.983	2	.05	.004	.162	.289	.320	2	.05	.002	.196	.452	.595
	.10	.005	.449	.798	.990		.10	.024	.311	.415	.462		.10	.004	.361	.618	.732

[1] $\alpha = 1$ for all items; $\delta = -2, -1.9, \ldots, 2, \delta \neq 0$; $\gamma = (0.4, 0.39, \ldots, 0.01)$; $\lambda = 0.8$ for all items

[2] $\alpha = 1$ for all items; $\delta = -2, -1.9, \ldots, 2, \delta \neq 0$; $\gamma = (0.4, 0.39, \ldots, 0.01)$; $\lambda = 0.9$ for $\delta_1(\ldots)\delta_{20}$; $\lambda = 0.8$ for $\delta_{21}(\ldots)\delta_{30}$; $\lambda = 0.7$ for $\delta_{31}(\ldots)\delta_{40}$

Except for $\theta = -2$ (both significance levels) and $\theta = -1$ (nominal 0.10 significance level), the detection rate was highest for the 1PLM; and except for $\theta = -2$ (both significance levels), the detection rate was lowest for the 4PLM with upper asymptotes $\lambda = 0.8$ for all items.

Results for $\mathcal{P}$ and ZU3, for 80 items. For 80-item tests and using $\mathcal{P}$, for all three models a substantially higher detection rate was found for each subdivision (Table 2). Using the 4PLM with $\lambda = 0.8$ for all items, the detection rate of ZU3 was rather sensitive to the closer spacing of the proportion-correct of the items. For example, at a nominal 0.05 significance level detection rates were 0.445 for the 4PLM (closer spacing), for $\theta = 0$; and 0.797 for the 1PLM (wider spacing), for $\theta = 0$. For the 4PLM with fixed λ's and $\theta = 0, 1, 2$, for the subdivision into eight subtest of 10 items the use of $\mathcal{P}$ yielded higher detection rates than ZU3. For the 4PLM with varying λ's and for $\theta = 1, 2$, detection rates were higher for $\mathcal{P}$. For $\theta = 0$, detection rates also were higher for $\mathcal{P}$ at nominal 0.10 significance level, but not at the 0.05 level.

Robustness results. For mild violations of IIO, detection rates (not tabulated) were similar (within a range of 0.01) to those discussed above. For example, for the 4PLM ($\lambda = 0.8$ for all items) and 40 items (eight subtests), for $\theta = 0$ the detection rate at nominal significance level of 0.05 was .198 compared to .193 (Table 1, middle panel) when IIO held. Our tentative conclusion is that the detection rates reported in the Tables 1 and 2 also apply when IIO is mildly violated.

Discussion

This study has presented new contributions to a relatively unexplored field in the analysis of item score patterns on a test. The concept of the PRF was studied extensively, and a new formal definition was given (Definition) that circumvents all problems with existing definitions. For finite test length, a discrete approximation of the PRF was given (Equation (17)), which can be estimated using (18). Deviations from the expected noninecreasingness of the PRF can be tested using a conservative cumulative hypergeometric test (Equation (21)). All developments presented here assume that the items have an IIO (Equation (5)).

Common person-fit methods provide a *scalar* value for each respondent, indicating which respondents have produced aberrant score patterns and which respondents have produced normal score patterns. Since the PRF is a *function*, it is a potentially powerful tool for the *diagnosis* of aberrance. Our simulation study showed that the detection rate for the local, cumulative hypergeometric $\mathcal{P}$ often was lower than that for the global, standard normal ZU3. This result does not discredit the PRF method: On the contrary, the ZU3 and the PRF methods provide different information, which may be used in two different ways.

First, when analyzing the score patterns on, for example, an achievement test a researcher may only be interested in detecting persons that have unusual score patterns on particular subsets of items because these patterns may be indicative of a particular kind of aberrant behavior. Examples may be cheating (unusually high scores on the difficult subsets) and lack of concentration (unusually low scores toward the end of the test). Thus, a researcher who has a priori expectations can use the PRF approach to test these hypotheses.

Second, the researcher may have no idea about the type of aberrance to be expected. First, a global person-fit statistic, such as ZU3, can be used for detecting aberrant item score patterns. Next, these aberrant patterns can be investigated using the PRF approach. Because the detection rate of the PRF method depends on the size of the item subset and because in an exploratory context the researcher may not know where the aberrance is to be expected, several sizes for item subsets should be tried.

Another approach to studying the decreasingness of the PRF may be to apply a *global* test, such as the Cochran-Armitage trend test (e.g., Agresti, 1990, pp. 100–102). When this test indicates that the PRF is *not* a decreasing function, the next step is to apply our *local*, cumulative hypergeometric test (Equation (21)). The Cochran-Armitage trend test tests the null hypothesis that a string of proportions does *not* decrease or increase against the alternative that the string decreases or increases. Acceptance of the null hypothesis means that the PRF is either a jumpy

TABLE 2.
Detection rate using the cumulative hypergeometric $\mathcal{P}$ and ZU3 for careless simulees on the 10 easiest items of the 80-item test. Detection rates are given only for $\mathbf{Y}_1 - \mathbf{Y}_2$ subtests

1PLM						4PLM[1]						4PLM[2]					
	Length subtest/#subtests						Length subtest/#subtests						Length subtest/#subtests				
θ	sig. lev.	40/2	20/4	10/8	ZU3	θ	sig. lev.	40/2	20/4	10/8	ZU3	θ	sig. lev.	40/2	20/4	10/8	ZU3
−2	.05	0	.004	.044	.158	−2	.05	0	.017	.131	.240	−2	.05	0	.015	.193	.330
	.10	0	.008	.083	.272		.10	0	.045	.176	.397		.10	0	.049	.343	.501
−1	.05	0	.018	.307	.470	−1	.05	0	.041	.323	.348	−1	.05	0	.041	.422	.483
	.10	0	.035	.487	.641		.10	0	.088	.504	.505		.10	0	.082	.607	.672
0	.05	0	.077	.663	.797	0	.05	0	.076	.496	.445	0	.05	0	.114	.663	.678
	.10	0	.145	.824	.889		.10	0	.155	.697	.607		.10	0	.211	.831	.809
1	.05	0	.374	.904	.971	1	.05	.002	.244	.637	.476	1	.05	0	.347	.807	.760
	.10	0	.549	.957	.981		.10	.003	.389	.795	.637		.10	0	.534	.910	.867
2	.05	.006	.719	.971	.997	2	.05	.010	.347	.670	.452	2	.05	.003	.565	.837	.818
	.10	.010	.851	.987	.998		.10	.039	.507	.823	.608		.10	.005	.701	.932	.897

[1] $\alpha = 1$ for all items; $\delta = -2, -1.95, \ldots, 2, \delta \neq 0$; $\gamma = (0.4, 0.395, \ldots, 0.005)$; $\lambda = 0.8$ for all items
[2] $\alpha = 1$ for all items; $\delta = -2, -1.95, \ldots, 2, \delta \neq 0$; $\gamma = (0.4, 0.395, \ldots, 0.005)$; $\lambda = 0.9$ for $\delta_1(\ldots)\delta_{40}$; $\lambda = 0.8$ for $\delta_{41}(\ldots)\delta_{60}$; $\lambda = 0.7$ for $\delta_{61}(\ldots)\delta_{80}$

curve or a horizontal line, and rejection means that the PRF is either decreasing or increasing, the appropriate option to be revealed by visual inspection. For example, let $J = 80$ and let eight 10-item subtests have fractions correct of *0.2*, *0.6*, 0.3, 0.3, 0.3, *0.1*, *0.2*, and *0.4* (italics indicate violations of expected nonincreasing ordering). The Cochran-Armitage trend test yielded $z^2 = 0.4082$, with df $= 1$, which obviously did not reject the null hypothesis. Thus, the conclusion is that the PRF is not a decreasing function and that there is evidence of aberrance. Current research addresses the usefulness of the Cochran-Armitage trend test in PRF research.

A specific person-fit statistic such as ZU3 was designed to detect aberrance, and evidence for this is obtained each time a respondent answers a relatively easy item incorrect while answering a more difficult item correct. Trend tests were not designed for detecting such peculiarities of the data. Since the PRF is a function for which a decreasing trend is expected, the idea of starting with a general trend test remains appealing, however. The design of a suitable global trend test could be one of the topics for future research.

References

Agresti, A. (1990). *Categorical data analysis*. New York: Wiley.

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

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

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

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

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.

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

Huynh, H. (1994). A new proof for monotone likelihood ratio for the sum of independent bernoulli random variables. *Psychometrika, 59*, 77–79.

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

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

Klauer, K.C., & 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.

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 Eduational Statistics, 4*, 269–290.

Lindgren, B.W. (1993). *Statistical theory*. New York: Chapman & Hall.

Lord, F.M. (1952). A theory of test scores. *Psychometrika Monograph No.7*.

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

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

Meijer, R.R. (Ed.). (1996). Person-fit research: Theory and applications [Special issue]. *Applied Measurement in Education, 9*(1).

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

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

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 statistics. *Psychometrika, 55*, 75–106.

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

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

Ramsay, J.O. (1995). A similarity-based smoothing approach to nondimensional item analysis. *Psychometrika, 60*, 323–339.

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

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

Rosenbaum, P.R. (1987b). Comparing item characteristic curves. *Psychometrika, 52,* 217–233.

Scheiblechner, H. (1995). Isotonic ordinal probabilistic models (ISOP). *Psychometrika, 60,* 281–304.

Sijtsma, K. (1998). Methodology Review: Nonparametric IRT approaches to the analysis of dichotomous item scores. *Applied Psychological Measurement, 22,* 3–31.

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., & Molenaar, I.W. (1987). Reliability of test scores in nonparametric item response theory. *Psychometrika, 52,* 79–97.

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

Suppes, P., & Zanotti, M. (1981). When are probabilistic explanations possible? *Synthese, 48,* 191–199.

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

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

Manuscript received 11 JUN 1998
Final version received 13 JUL 1999