Comparing simulated and theoretical sampling distributions of the U3 person-fit statistic

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

Published in:
Applied Psychological Measurement

Publication date:
2002

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

[Link to publication in Tilburg University Research Portal](#)

Citation for published version (APA):
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(1), 88-108.

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Applied Psychological Measurement

<http://apm.sagepub.com>

Comparing Simulated and Theoretical Sampling Distributions of the U3 Person-Fit Statistic

Wilco H. M. Emons, Rob R. Meijer and Klaas Sijtsma

Applied Psychological Measurement 2002; 26; 88

DOI: 10.1177/0146621602026001006

The online version of this article can be found at:

<http://apm.sagepub.com/cgi/content/abstract/26/1/88>

Published by:

SAGE Publications

<http://www.sagepublications.com>

Additional services and information for *Applied Psychological Measurement* can be found at:

Email Alerts: <http://apm.sagepub.com/cgi/alerts>

Subscriptions: <http://apm.sagepub.com/subscriptions>

Reprints: <http://www.sagepub.com/journalsReprints.nav>

Permissions: <http://www.sagepub.com/journalsPermissions.nav>

Citations (this article cites 19 articles hosted on the
SAGE Journals Online and HighWire Press platforms):

<http://apm.sagepub.com/cgi/content/refs/26/1/88>

Comparing Simulated and Theoretical Sampling Distributions of the $U3$ Person-Fit Statistic

Wilco H. M. Emons, Tilburg University

Rob R. Meijer, University of Twente

Klaas Sijtsma, Tilburg University

The accuracy with which the theoretical sampling distribution of van der Flier's person-fit statistic $U3$ approaches the empirical $U3$ sampling distribution is affected by the item discrimination. A simulation study showed that for tests with a moderate or a strong mean item discrimination, the Type I error rates were either too high or too low to be used in practice. It was concluded that the use of standard normal deviates

for the standardized version of the $U3$ statistic may be problematic. Nevertheless, the $U3$ statistic is suitable for evaluating the relative likelihood of item score patterns, for example, if one wishes to select a fixed percentage of the most improbable patterns. *Index terms: item response theory, nonparametric item response models, person-fit analysis, $U3$ person-fit statistic, response aberrance.*

Person fit is concerned with the detection of item score patterns that have a low probability given what is expected under a particular test model or given the majority of item score patterns in the sample. Unusual item score patterns should be detected because they may not give an adequate description of the respondent's trait level. As a consequence, the validity of the individual test scores may be affected (Meijer, 1997, 1998; Schmitt, Chan, Sacco, McFarland, & Jennings, 1999). Examples of aberrant response behavior include cheating, guessing, plodding, and extreme creativity (Meijer, 1994). Person-fit indices have been used to identify schools that have curricula that did not match test content (Harnisch & Linn, 1981) and to identify students with certain language deficiencies on an intelligence test (van der Flier, 1980). Recently, Meijer and Sijtsma (1995, 2001) provided reviews of methods for evaluating the fit of item score patterns.

In parametric item response theory (IRT), the relationship between the latent trait θ and the item score is described by a parametric item response function (IRF). Several person-fit studies used statistics that were formulated in the context of parametric IRT (Levine & Rubin, 1979; Drasgow, Levine, & McLaughlin, 1987) to evaluate the likelihood of item score patterns on an individual level. Attempts to formulate person-fit analysis outside the context of parametric IRT yielded statistics that compare an individual's item score pattern with the item score patterns of the other persons in the group.

This study deals with person-fit analysis in the context of nonparametric IRT (NIRT; Mokken & Lewis, 1982; Sijtsma, 1998). Unlike parametric IRT models, NIRT models do not assume a particular parametric form for the IRF. A typical assumption of a NIRT model is that the IRF is a nondecreasing function of θ . Given this constraint, any form of the IRF is acceptable. NIRT models imply ordinal measurement of persons or items on a latent trait θ (Hemker, Sijtsma, Molenaar, & Junker, 1997).

These models can be useful for the analysis of test data, especially when an ordering of respondents on θ is sufficient for the application envisaged (Sijtsma, 1998).

In nonparametric person-fit analysis, an item score pattern is considered nonfitting if it is improbable given a NIRT model (Meijer & Sijtsma, 1995). Several nonparametric or group-based statistics have been proposed (see, e.g., Rudner, 1983; Meijer & Sijtsma, 2001). For most of these statistics, the null distribution of statistics given normal response behavior under the NIRT model is unknown. As a result, it cannot be decided on the basis of significance probabilities whether a score pattern is unlikely given a nominal Type I error rate. Alternatively, rules of thumb for classifying item score patterns were proposed, which were based on simulated data or on a limited number of empirical data sets (e.g., such rules were proposed for the H^T coefficient [Sijtsma & Meijer, 1992] and the C index [Harnisch & Linn, 1981]). Often, it is difficult to generalize these rules of thumb to other data sets.

The $U3$ statistic (van der Flier, 1980, 1982), however, is a group-based statistic with a known null distribution. This sampling distribution can be used to obtain critical values for classifying item score patterns as fitting or nonfitting. Furthermore, $U3$ conditional on the number-correct score is monotonically related to the significance probability (van der Flier, 1980, p. 61). Some research has been done with $U3$ (e.g., Meijer, Molenaar, & Sijtsma, 1994), which showed high detection rates for nonfitting item score patterns, in particular for long tests and items with high discrimination power. These detection rates were studied using samples with a known mixture of fitting and nonfitting item score patterns. In real test applications, researchers usually have little or no knowledge about the percentage of respondents in the sample who produced a nonfitting pattern and, hence, a sampling distribution is needed for hypothesis testing (Molenaar & Hoijtink, 1990).

This study extended the work of van der Flier (1980); van der Flier (1980) found that for tests with at least 29 items, the means and standard deviations of the conditional $U3$ sampling distributions based on simulated data were closely approximated by the theoretically derived means and standard deviations. Comparison of the simulated cumulative distribution of $U3$ with the theoretical approximation of the cumulative sampling distribution showed differences of at least 0.06 on the vertical probability scale. These comparisons, however, were based on sampling distributions that were simulated under IRT models that assume horizontal IRFs, which are rather unrealistic. It would be interesting also to simulate sampling distributions using more realistic sets of IRFs and compare the results to van der Flier's results. The purpose of this study was to investigate whether the theoretical sampling distribution of $U3$ is in agreement with simulated sampling distribution under different IRT models, and varying test and item characteristics. In particular, we investigated the usefulness of critical values based on the theoretical sampling distribution of $U3$ for classifying item score patterns.

The $U3$ Statistic

We assume that a test consists of J dichotomously scored items. Let X_j ($j = 1, \dots, J$) be the random variable for the binary item scores, with the value 1 for a correct (or coded) response and 0 otherwise. Also, $\mathbf{X} = (X_1, \dots, X_J)$ is the random vector of the item score variables. Furthermore, let X_+ be the random variable for the unweighted sum score, $X_+ = \sum_{j=1}^J X_j$. Let π_j ($j = 1, \dots, J$) be the proportion of correct responses to item j in the population and let its sample estimate be denoted by $\hat{\pi}_j$. Throughout this study, it will be assumed that the items are ordered from easy to difficult; that is, $\pi_1 \geq \pi_2 \geq \dots \geq \pi_J$.

An item score pattern with correct responses in the first X_+ positions and incorrect responses in the remaining $J - X_+$ positions is called a *Guttman pattern* because it meets the requirement of the Guttman (1950) scalogram. Analogously, a pattern with all correct responses in the last X_+ positions and incorrect responses in the remaining positions is called a *reversed Guttman pattern*.

The $U3$ statistic for the pattern $\mathbf{X}$ that yields X_+ items correct is given by

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

For fixed X_+ , all terms are constant except

$$W(\mathbf{X}) \equiv \sum_{j=1}^J X_j \log\left(\frac{\pi_j}{1-\pi_j}\right), \quad (2)$$

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

Van der Flier (1980, 1982) derived the expected value and the variance of $W(\mathbf{X})$ across replications, given fixed X_+ . If the number of items is at least 20, and the π values show reasonable variance (van der Flier, 1980, p. 295; the author does not quantify what he considers reasonable), then $W(\mathbf{X})$ given X_+ is normally distributed, with mean and variance

$$\mu_{[W(\mathbf{X})|X_+]} = \sum_{j=1}^J \pi_j \log\left(\frac{\pi_j}{1-\pi_j}\right) + \frac{\sum_{j=1}^J \pi_j(1-\pi_j) \log\left(\frac{\pi_j}{1-\pi_j}\right)}{\sum_{j=1}^J \pi_j(1-\pi_j)} \left(X_+ - \sum_{j=1}^J \pi_j\right) \quad (3)$$

and

$$\sigma_{[W(\mathbf{X})|X_+]}^2 = \sum_{j=1}^J \pi_j(1-\pi_j) \left[\log\left(\frac{\pi_j}{1-\pi_j}\right)\right]^2 - \frac{\left[\sum_{j=1}^J \pi_j(1-\pi_j) \log\left(\frac{\pi_j}{1-\pi_j}\right)\right]^2}{\sum_{j=1}^J \pi_j(1-\pi_j)}, \quad (4)$$

respectively (van der Flier, 1980, p. 66).

Consequently, $U3$ is normally distributed with conditional expectation and conditional variance

$$E(U3|X_+) = \frac{\sum_{j=1}^{X_+} \log\left(\frac{\pi_j}{1-\pi_j}\right) - \mu_{[W(\mathbf{X})|X_+]}}{\sum_{j=1}^{X_+} \log\left(\frac{\pi_j}{1-\pi_j}\right) - \sum_{j=J-X_++1}^J \log\left(\frac{\pi_j}{1-\pi_j}\right)} \quad (5)$$

and

$$Var(U3|X_+) = \frac{\sigma_{[W(\mathbf{X})|X_+]}^2}{\left[\sum_{j=1}^{X_+} \log\left(\frac{\pi_j}{1-\pi_j}\right) - \sum_{j=J-X_++1}^J \log\left(\frac{\pi_j}{1-\pi_j}\right)\right]^2}, \quad (6)$$

respectively. The standardized version, denoted $ZU3$, is asymptotically standard normally distributed. The value of $ZU3$ for $\mathbf{X}$ yielding X_+ items correct is obtained using Equations 5 and 6,

$$ZU3(\mathbf{X}) = \frac{U3(\mathbf{X}) - E(U3|X_+)}{\sqrt{Var(U3|X_+)}}. \quad (7)$$

For a comprehensive discussion of the derivation of Equations 1, 3, and 4, see van der Flier (1980, pp. 62-67).

Next, we discuss the assumptions that were made in the derivation of $U3$. The derivation of the theoretical mean and variance of $U3$ consisted of two steps. In the first step, the unconditional expectation and variance of $W(\mathbf{X})$ in the population were derived. Then, in the second step, the conditional distribution of $W(\mathbf{X})$ given X_+ was derived from the bivariate distribution of X_+ and $W(\mathbf{X})$.

In the first step, the unconditional distributions of $W(\mathbf{X})$ and X_+ were obtained by assuming that the item scores are statistically independent in the *population*, which implies that for two arbitrarily chosen items, say j and j^* , in the population $Cov(X_j, X_{j^*}) = 0$. In IRT, the assumption of statistical independence between item scores holds if either the variance of θ equals 0 or the items have IRFs that are constant functions of θ . Flat IRFs imply that the items are unrelated to θ , which means that the items do not discriminate between respondents. Thus, differences between observed scores are entirely due to measurement error and, therefore, represent unsuccessful measurement. In practice, item constructors select those items from a set of candidate items that have high discrimination power because these are the most informative items. Such items produce high positive covariance between the item scores (Mokken, 1971, p. 131; Sijtsma, 1998).

In the second step, it was assumed that $W(\mathbf{X})$ and X_+ follow a bivariate normal distribution, with unconditional univariate distributions $W(\mathbf{X})$ and X_+ equal to the estimated distributions obtained in the first step. Given the bivariate normally distributed random vector $(W(\mathbf{X}), X_+)$, the conditional mean and variance of $W(\mathbf{X})$ given X_+ were found by Equation 3, which is a linear regression function of $W(\mathbf{X})$ on X_+ , and Equation 4 (van der Flier, 1980, pp. 65-66; see also Lindgren, 1993, pp. 423-425). It is unknown, however, to what extent the nonzero covariances between item scores affect the bivariate normal distribution of X_+ and $W(\mathbf{X})$ and, consequently, affect the feasibility of using Equations 3 and 4 to estimate the conditional distribution of $W(\mathbf{X})$.

Given that in practice the assumption of statistical independence between item scores is unrealistic, the theoretical distribution of $ZU3$ may not be valid in practical applications. Given this uncertainty, in this study we investigated whether $ZU3$ follows a standard normal distribution when the data have a more realistic IRT structure.

Method

Design

Data were simulated under a design with four independent factors. The first factor was the IRT model used for simulating data. Four different IRT models were used. The first level was a unidimensional, locally independent IRT model with flat IRFs; that is, the probability of giving a correct answer on item j , $P_j(\theta)$, is a constant function of θ : $P_j(\theta) = P_j$. This model implies that the covariance between two arbitrarily chosen items is 0. We refer to this model as the Model of Marginal Independence (MMI). Obviously, the MMI with its flat IRFs is an unrealistic model, but it was included in this study because this is the model under which van der Flier (1980) derived the theoretical distribution of $U3$. Not only were we interested to know whether the empirical distribution matched the theoretical distribution under the MMI, but the MMI also served as a benchmark for simulations

under other, more realistic IRT models, which did *not* underlie the theoretical distribution properties of $U3$.

The second and third levels were two unidimensional, locally independent parametric IRT models: the restrictive Rasch model (RM; Rasch, 1960) and the more liberal three-parameter logistic model (3PLM; Birnbaum, 1968). Following Hambleton and Swaminathan (1985, p. 47), the RM can be written as

$$P_j(\theta) = \frac{\exp[\bar{\alpha}(\theta - \delta_j)]}{1 + \exp[\bar{\alpha}(\theta - \delta_j)]},$$

where $\bar{\alpha}$ is the common level of discrimination for all J items in the test and δ_j is a location parameter. Hambleton and Swaminathan (1985, p. 47) noted that the RM can also be written with $\bar{\alpha}$ incorporated into the θ scale, by rescaling $\theta' = \bar{\alpha}\theta$ and $\delta' = \bar{\alpha}\delta$. Thus, although authors often choose to write the RM with $\bar{\alpha} = 1$, in fact all the RM assumes is that $\bar{\alpha}$ is the same for all J items, and $\bar{\alpha} = 1$ can always be obtained by an appropriate rescaling of θ .

The 3PLM is defined as

$$P_j(\theta) = \gamma_j + (1 - \gamma_j) \frac{\exp[\alpha_j(\theta - \delta_j)]}{1 + \exp[\alpha_j(\theta - \delta_j)]},$$

where γ_j is the lower asymptote for $\theta \rightarrow -\infty$ and α_j is monotonically related to the slope of the IRF at location δ_j . This slope parameter is also known as the discrimination parameter. The fourth level was Mokken's Monotone Homogeneity Model (MHM; Mokken, 1971), which is a unidimensional, locally independent nonparametric IRT model. The MHM assumes that the IRFs are monotonely nondecreasing functions; that is, $P_j(\theta_a) \leq P_j(\theta_b)$, whenever $\theta_a < \theta_b$. The MHM is the most liberal of the IRT models investigated in this study.

The second factor was item discrimination, which directly affects the covariance between the items: the higher the discrimination power, the higher the covariance (Hemker, Sijtsma, & Molenaar, 1995). Three levels of item discrimination were studied: weak, moderate, and strong, to be defined shortly. The third factor was test length, with three levels: $J = 20, 40$, and 80 . Finally, the fourth factor was the spread of the item difficulties. Two levels were studied: small and large, to be defined shortly.

The RM, the 3PLM, and the MHM were completely crossed with the three levels of item discrimination, the three levels of test length, and the two levels of spread of item difficulty. For the MMI were fully crossed with the three levels of test length and the two levels of spread of item difficulty. The result is two cross-factorial designs with $3 \times 3 \times 3 \times 2 = 54$ cells and $3 \times 2 = 6$ cells, respectively.

Data Simulation and Specification of the Factors

Model of marginal independence. Data matrices for the MMI were simulated as follows. For each level of test length, two sets of P_j s were specified. One set had P_j s equidistant on the interval $[0.30, 0.70]$, corresponding to small spread of item difficulties. The other set had P_j s equidistant on the interval $[0.10, 0.90]$, corresponding to large spread of item difficulties. The item scores were simulated by drawing a random number y from the uniform distribution on the interval $[0, 1]$; when $y \leq P_j$, the item score was 1, and 0 otherwise.

Parametric IRT models. For both the RM and the 3PLM, the δ s were chosen to be equidistant on the interval $[-1, 1]$ for the condition with small spread, and equidistant on the interval $[-2, 2]$ for the condition with large spread. For the RM, increasing the spread of the θ distribution has the effect of increasing the discrimination power of all items. Mathematically, this is identical to increasing $\bar{\alpha}$ for all J items, while holding the spread of θ constant. Therefore, under a standard normal distribution

for θ , we chose $\bar{\alpha} = 0.5, 1$, and 2 for weak, moderate, and strong discrimination, respectively. For the 3PLM, each set of J items had item discrimination parameters that for separate sets were sampled from one of the following truncated normal distributions: $\alpha \sim N(0.5, 0.25)$, truncated at $(0, 1)$ (Weak); $\alpha \sim N(1, 0.25)$, truncated at $(0.5, 2)$ (Moderate); and $\alpha \sim N(2, 0.25)$, truncated at $(1.0, 3.0)$ (Strong). Moreover, for each test the γ 's were sampled from a uniform distribution on the interval $[0, 0.2]$. For the RM and the 3PLM, the item scores were simulated by drawing a random number y from the uniform distribution on the interval $[0, 1]$; when $y \leq P_j(\theta)$, the item score was 1, and 0 otherwise.

Mokken's monotone homogeneity model. For the MMI and the parametric models, the conditional probabilities, $P_j(\theta)$, were used for simulating item scores. However, the MHM does not parametrically define the IRFs and, consequently, numerical values for the conditional success probabilities cannot be obtained in an obvious way. Most simulation studies in the context of NIRT used parametric IRT models to generate the data matrices (for example, Meijer et al. [1994] used the 2PLM, and Hemker et al. [1995] used the graded response model for simulating polytomous item scores). The choice of logistic IRFs may somewhat limit the generalizability of the results. Alternatively, we used a procedure for simulating data that only used the feature of monotonely nondecreasing IRFs, without any restrictions on the functional form.

The following procedure was used to simulate data under the MHM. For different data sets, the procedure used Mokken's (1971, p. 185) definitions of a weak scale, a medium scale, and a strong scale. These definitions use the item scalability coefficient H_j (Mokken, 1971, p. 152; Mokken & Lewis, 1982), which is defined using the π_j 's and the bivariate proportions of having items j and k correct, denoted π_{jk} :

$$H_j = \frac{\sum_{k \neq j} (\pi_{jk} - \pi_j \pi_k)}{\sum_{k \neq j} \pi_j (1 - \pi_k)}, \text{ with } \pi_j \leq \pi_k, \quad (8)$$

and the overall scalability coefficient, denoted H , which is a positively weighted sum of the J H_j 's (Mokken, 1971, p. 151). A weak scale is a set of J items that (a) have positive covariances, (b) each has an item scalability coefficient $H_j \geq c$ (in practice, it is recommended to set c equal to 0.30), and (c) together have an overall scalability coefficient $0.30 \leq H < 0.40$. The first requirement is a necessary condition for all unidimensional, locally independent IRT models with monotonely nondecreasing IRFs (Holland & Rosenbaum, 1986; Junker, 1993), and the second requirement stipulates that items have at least weak discrimination (Sijtsma, 1998). The third requirement expresses the degree of scalability corresponding to a weak scale. A medium scale differs from a weak scale in that $0.40 \leq H < 0.50$. The items from a medium scale tend to have moderate discrimination. A strong scale has $H \geq 0.5$. The items from this scale tend to have strong discrimination. It may be noted that both the scalability coefficients H_j and H depend heavily on the distribution of the person parameters (Hemker et al., 1995).

For each level of test length and spread of item difficulties, for a given distribution of θ , sets of IRFs were defined to constitute either a weak scale, a medium scale, or a strong scale. The procedure defined each IRF by 13 discrete points, which were connected by straight lines. Each of these points was defined by coordinates $(\theta_{jt}, P_j(\theta_{jt}))$, with $t = 1, \dots, 13$. Consecutive probabilities satisfied the inequality restriction $P_j(\theta_{jt}) \leq P_j(\theta_{j,t+1})$, whenever $\theta_{jt} < \theta_{j,t+1}$ (see Figure 1). The success probability for a fixed θ_i was obtained by means of linear interpolation: If $\theta_{jt} \leq \theta_i \leq \theta_{j,t+1}$, then

$$P_j(\theta_i) = P_j(\theta_{jt}) + \frac{\theta_i - \theta_{jt}}{\theta_{j,t+1} - \theta_{jt}} [P_j(\theta_{j,t+1}) - P_j(\theta_{jt})]. \quad (9)$$

Figure 1
Fragments of Three Item Response Functions on $\theta = (-1.5, 1.5)$
Used for Simulation Under the Monotone Homogeneity Model

Next, we discuss the choice of the $P_j(\theta_{jt})$ s and the θ_{jt} s. First, for each item j , the values for $P_j(\theta_{jt})$, $t = 1, \dots, 13$, were generated: $P_j(\theta_{j1})$, $P_j(\theta_{j7})$, and $P_j(\theta_{j13})$ were a priori fixed at 0.0, 0.5, and 1.0, respectively. The remaining 10 values for $P_j(\theta_{jt})$ were sampled from a uniform distribution in such a way that (a) the IRFs were monotonically nondecreasing and (b) some IRFs approached the values of 0 and 1 reasonably slowly, whereas others were much steeper. For example, first we drew $P_j(\theta_{j10})$ from a uniform distribution between 0.5 and 1. The next value that was drawn was $P_j(\theta_{j8})$. Given the value of $P_j(\theta_{j10})$, $P_j(\theta_{j8})$ was drawn from the uniform distribution on the interval $[0.5, P_j(\theta_{j10})]$. This procedure indeed produced IRFs that are monotonely nondecreasing, in some cases IRFs with steep slopes and in other cases IRFs with flatter or much flatter slopes.

Second, the corresponding values of θ_{jt} were specified as follows. The J values of θ_{j7} were fixed equidistant on the interval $(-0.5, 0.5)$ (small spread of item difficulties) or $(-1.25, 1.25)$ (large spread of item difficulties), thus specifying the location at the latent scale for which $P_j(\theta_{j7}) = 0.5$. In addition, for each item j , θ_{j1} and θ_{j13} were fixed at -9 and 9 , respectively, thus specifying the location at which the IRFs reached either their minimum or maximum.

Next, the remaining values of θ_{jt} were specified. Let Δ denote the constant distance between two consecutive values of θ_{jt} for item j : $\Delta = \theta_{jt+1} - \theta_{jt}$, $t = 2, \dots, 11$. Because fixed θ_{j7} and Δ together implied the other values of θ_{jt} , and θ_{j7} was already chosen, we had to specify Δ for all j to define our IRFs. For a set of IRFs to constitute a weak scale, Δ had to be chosen such that, in combination with the distribution of θ , the resulting IRFs yielded an overall scalability H value between 0.30 and 0.40. This was done as follows. First, we assumed a standard normal θ . Second, we chose an initial value for Δ , say $\Delta^{(o)} = 0.3$. This choice determined the shape of the IRFs and together with the standard normal θ , the π_{js} s and the π_{jks} s could be determined. These values were inserted in Equation 8, and the H_j s were calculated. Furthermore, H was calculated as a weighted average of the H_j s. When the overall H was not in the interval for weak scales, other values for Δ were tried iteratively until a satisfactory H was found.

Given the resulting set of weak-scale IRFs, moderate and strong scales were constructed by gradually increasing the variance of θ , because for fixed IRFs, a higher θ variance has the effect of producing higher discrimination power (Hemker et al., 1995; also, Roskam, van den Wollenberg, & Jansen, 1986; Mokken, Lewis, & Sijtsma, 1986). The values for the θ variance that were used

were 1.3 (moderate discrimination) and 2.0 (strong discrimination), producing H values in the population of $0.40 \leq H \leq 0.50$, and $H > 0.50$, respectively.

Again, item scores were simulated by randomly drawing a y from the uniform distribution on the interval $[0, 1]$; when $y \leq P_j(\theta)$, the item score was 1, and 0 otherwise, with randomly drawn θ and $P_j(\theta)$ calculated using linear interpolation as in Equation 9.

Calibration

For each cell in the design, a separate calibration sample was simulated to obtain sample values of the item difficulties, the $\hat{\pi}$ s, given the postulated IRT model and a sample from the theoretical θ distribution. More specifically, it may be noted that at the population level,

$$\pi_j = \int P_j(\theta) dG(\theta),$$

where $G(\theta)$ is a cumulative distribution function. For a specific choice of $P_j(\theta)$ and a sample of 5,000 θ s, numerical integration was used for calculating the fractions $\hat{\pi}_j$ ($j = 1, \dots, J$).

These $\hat{\pi}$ values were needed for determining the ordering of the items according to their difficulty and also for calculating the theoretical expressions (Equations 3–6) for the mean and the standard deviation of the conditional distribution of $U3$, which in turn were needed for calculating $ZU3$ (Equation 7).

Simulating the Distribution of $U3$

At each level of X_+ , 1,000 item score patterns were simulated. Because $ZU3$ is not defined for $X_+ = 0$ and $X_+ = J$, conditional distributions for these total scores could not be calculated. Next, values of $ZU3$ were computed for each item score pattern. The 1,000 $ZU3$ values were used to obtain the empirical distribution of $ZU3$ at score level $X_+ = 1, \dots, J - 1$, respectively. The simulated conditional sampling distributions of $ZU3$ were evaluated by examining the first four moments.

To study whether the normal approximation held in the tails of the distribution, simulated Type I error rates (false alarms) were studied at three significance levels: .10, .05, and .01. Because large values of $ZU3$ indicate nonfitting item score patterns, the significance probabilities (one-side tests) in the right tail of the sampling distribution were of interest. Critical standard normal deviates corresponding to the three significance levels were 1.28, 1.65, and 2.33, respectively. Thus, item score patterns yielding a $ZU3$ score that exceeded a particular critical value were classified as nonfitting. Because only fitting item score patterns were simulated, each pattern classified as nonfitting was incorrectly classified.

Results

Appropriateness of the Theoretical Mean and Standard Deviation of $U3$

First, we consider the estimates of the theoretical conditional mean of $U3$. The estimates of the theoretical conditional mean of $U3$ for low number-correct score X_+ or high X_+ were negative for the three levels of test length and the two levels of spread of item difficulties. For example, Table 1 gives the estimates of the theoretical mean and standard deviation of $U3$ at $X_+ \leq 5$ and $X_+ \geq 34$, obtained for the 40-item test under the MMI, with π values equidistant on the interval $[0.10, 0.90]$. For these X_+ levels, the estimated mean of $U3$ was negative and large given the standard deviation. Note that, theoretically, a negative mean value of $U3$ is impossible because $0 \leq U3 \leq 1$ by definition. Thus, the simulated sampling distributions of $U3$ and $ZU3$ at low or high score levels are biased.

Table 1
Theoretical Mean and Standard
Deviation (*SD*) of *ZU3* Under
the Model of Marginal Indepen-
dence, $J = 40$, and Large
Spread of Item Difficulties

X_+	Mean	<i>SD</i>
1	-1.793	0.363
2	-0.709	0.333
3	-0.348	0.232
4	-0.167	0.182
5	-0.059	0.152
...	...	...
34	-0.015	0.132
35	-0.095	0.152
36	-0.216	0.182
37	-0.418	0.232
38	-0.823	0.333
39	-2.040	0.636

Evaluation of the First Four Moments of the Simulated Distribution of *ZU3*

Model of marginal independence. Figure 2 shows the conditional means and conditional variances of *ZU3* simulated under the MMI as a function of X_+ , based on $J = 80$, for the two levels of spread of the π_j s. The two lower curves show the simulated means, and the two higher curves show the simulated standard deviations. Because under the MMI the probability of simulating a pattern with $X_+ < 29$ or $X_+ > 52$ was approximately equal to zero, no results were obtained for these levels of X_+ . In Figure 2, it can be seen that the simulated conditional mean of *ZU3* for score groups with $30 \leq X_+ \leq 48$ (small spread of π_j s) and $35 \leq X_+ \leq 46$ (large spread of π_j s) were close to the expected value of 0; deviations were found to be smaller than 0.10. In the tails of the X_+ distribution, the conditional means of *ZU3* were larger than expected. It can also be seen that for both levels of spread of π_j s, and all levels of X_+ , the simulated variances of *ZU3* were close to the expected value of 1; deviations were smaller than 0.095 for small spread of the π_j s, and smaller than 0.054 for large spread of the π_j s. Similar results were found for $J = 20$ and $J = 40$ (results not shown here).

Results for parametric and nonparametric IRT models. For the 3PLM, Figure 3 shows the simulated conditional mean of *ZU3* based on $J = 40$ and small spread of the item difficulties (δ s), for the three levels of item discrimination. Figure 3 shows that the simulated conditional mean of *ZU3* has a curvilinear relation with X_+ ; *ZU3* was larger than the expected value of 0 in the tails of the X_+ distribution, and smaller for X_+ values in the center. Furthermore, Figure 3 shows that the difference between the simulated conditional mean value of *ZU3* and the theoretical mean value of 0 increased as item discrimination increased. The simulated conditional mean of *ZU3* ranged from -0.169 to 0.728 (weak discrimination), -0.331 to 1.166 (moderate discrimination), and -1.004 to 2.530 (strong discrimination). Comparable results were found for $J = 20$ and $J = 80$ (not shown here).

Figure 4 shows the simulated conditional variance of *ZU3* based on the 3PLM, for $J = 40$ and small spread of the δ s, for the three levels of item discrimination. It can be seen that in the middle range of X_+ , the conditional variance was closest to the expected value of 1, whereas in the tails, the variance was smaller than 1. Furthermore, for all levels of X_+ , the distribution of *ZU3* was

Figure 2
Conditional Mean and Standard Deviation (SD) of $ZU3$ Simulated Under the Model of Marginal Independence, With $J = 80$, for Two Levels of Spread of Item Difficulty (Small and Large) (1,000 Observations at Each Level of Sum Score)

Figure 3
Simulated Conditional Mean of $ZU3$ (Three-Parameter Logistic Model), With $J = 40$ and Small Spread of Item Difficulties, for Three Levels of Item Discrimination (Weak, Moderate, and Strong) (1,000 Observations at Each Level of Sum Score)

positively skewed; for weak item discrimination, the skewness varied from -0.039 to 0.386 , and for strong item discrimination, the skewness varied from 0.198 to 0.829 . No particular trends were found for the kurtosis (no results tabulated for skewness and kurtosis).

Similar results were found for the other levels of J , and also for the simulations under the RM and the MHM. Thus, in general, as item discrimination increased, simulated distributions showed means that were more biased, had smaller variance, and had larger skewness, but kurtosis was not affected.

Figure 4
Simulated Conditional Standard Deviation of $ZU3$ (Three-Parameter Logistic Model),
With $J = 40$ and Small Spread of Item Difficulties, for Three Levels of Item Discrimination (Weak, Moderate, and Strong) (1,000 Observations at Each Level of Sum Score)

Evaluation of Simulated Type I Error Rates

Results for the model of marginal independence. Table 2 shows the simulated Type I error rates at three significance levels for $J = 20, 40$, and 80 , across two levels of spread of the π_j s. No results were obtained for relatively low X_+ or high X_+ because under the MMI these values had approximately zero probability. Moreover, to avoid cumbersome tables, only the Type I error rates at every second X_+ are given. Table 2 shows that for each level of test length, the simulated Type I error rates in the middle of the X_+ distribution were close to the nominal Type I error rates; differences were smaller than 0.02 . In the tails of the X_+ distribution, Type I error rates were 1.5 times higher than expected. Higher Type I error rates were found in the tails of the X_+ distribution because simulated $ZU3$ distributions were shifted to the right relative to the expected distribution. Furthermore, it can be seen from Table 2 that for large spread of the π_j s, the simulated Type I error rates increased more rapidly for X_+ near the tails of the distribution.

The simulated Type I error rates reported in Table 2 indicate that if the MMI holds, for practical purposes the theoretical distribution may be useful to investigate nonfitting item score patterns, except for those patterns with high or low X_+ . However, it may be noted that the MMI is only relevant from a theoretical point of view, but not in practical applications of IRT.

Results for parametric IRT models. For the 3PLM, in Tables 3 through 5 the simulated Type I error rates at the three significance levels are given for the three levels of item discrimination, for the two levels of spread of item difficulties, and for $J = 20, 40$, and 80 . Because the simulated Type I error rates did not substantially vary between two adjacent levels of X_+ , and because we wanted to avoid cumbersome tables, only the Type I error rates at every fourth X_+ (for $J = 20$ and $J = 40$) or fifth X_+ (for $J = 80$) are given.

Tables 3 through 5 show that for weak discrimination and X_+ in the middle of the X_+ distribution, the simulated Type I errors were only slightly different from the expected Type I error rates. For example, for $J = 80$, weak discrimination, small spread of the δ s, and a significance level of $.10$, the Type I error rates for $30 \leq X_+ \leq 70$ varied from 0.110 to 0.123 (Table 5), and the largest difference between the expected and simulated Type I error rate was 0.043 at $X_+ = 51$ (not shown).

Table 2
Simulated Type I Error Rates (Model of Marginal Independence) at Three Significance Levels (Sign. Lev.) as Function of X_+ , for Two Levels of Spread of Item Difficulties and Three Levels of Test Length (1,000 Observations at Each Score Level)

X_+	Spread of Item Difficulties					
	Small			Large		
	Sign. Lev.			Sign. Lev.		
	.10	.05	.01	.10	.05	.01
<i>J</i> = 20						
4	.153	.068	.014	.289	.174	.058
6	.125	.066	.019	.152	.094	.032
8	.097	.061	.017	.116	.064	.012
10	.107	.062	.015	.084	.045	.012
12	.093	.043	.010	.125	.075	.021
14	.122	.066	.011	.138	.086	.022
16	.140	.071	.010	.314	.199	.058
<i>J</i> = 40						
12	.148	.076	.018	.190	.111	.028
14	.120	.067	.014	.137	.075	.022
16	.098	.049	.008	.092	.051	.015
18	.079	.032	.004	.104	.058	.015
20	.108	.054	.011	.101	.050	.019
22	.089	.041	.012	.115	.063	.015
24	.110	.068	.013	.132	.070	.019
26	.130	.064	.012	.168	.097	.018
28	.127	.062	.008	.218	.133	.039
<i>J</i> = 80						
30	.114	.054	.007	.157	.095	.031
32	.086	.048	.007	.116	.066	.018
34	.099	.054	.006	.143	.080	.012
36	.111	.050	.010	.090	.050	.014
38	.106	.049	.011	.112	.066	.016
40	.092	.050	.010	.090	.048	.015
42	.102	.055	.009	.090	.049	.017
44	.107	.059	.016	.099	.052	.012
46	.133	.059	.013	.123	.059	.018
48	.132	.059	.010	.119	.063	.017
50	.119	.069	.015	.133	.075	.022
51	.116	.061	.019	.156	.096	.029

in the table). It may be noted that no results were obtained for very low X_+ or very high X_+ because under weak item discrimination these values had approximately zero probability.

For moderate and strong item discrimination, differences between simulated and expected Type I error rates increased substantially, and the simulated Type I error rates were much smaller than expected for X_+ in the middle range, and much larger than expected for X_+ in the tails. For example, for $J = 80$, small spread of the δ s, and strong item discrimination, Table 5 shows that for all three significance levels, the simulated Type I error rates were smaller than 0.02 for $30 \leq X_+ \leq 60$. Furthermore, in the tails of the X_+ distribution, significance levels were found that were 5 times

Table 3
Simulated Type I Error Rates (Three-Parameter Logistic Model)
at Three Significance Levels (Sign. Lev.), for $J = 20$, Three
Different Levels of Item Discrimination, and Two Levels of Spread
of Item Difficulties (1,000 Observations at Each Score Level)

X_+	Item Discrimination								
	Weak			Moderate			Strong		
	Sign. Lev.			Sign. Lev.			Sign. Lev.		
	.10	.05	.01	.10	.05	.01	.10	.05	.01
Small Spread of Item Difficulties									
1	.018	.018	.000	.465	.207	.000	.649	.463	.059
4	.105	.037	.006	.231	.127	.028	.260	.152	.053
8	.112	.063	.012	.105	.063	.010	.064	.036	.010
12	.106	.058	.016	.070	.036	.010	.025	.011	.003
16	.094	.044	.005	.062	.022	.002	.031	.007	.001
19	.013	.000	.000	.034	.012	.000	.096	.021	.000
Large Spread of Item Difficulties									
1	.463	.169	.009	1.000	1.000	.377	1.000	1.000	1.000
4	.184	.102	.014	.459	.328	.142	.657	.535	.313
8	.094	.055	.010	.133	.075	.031	.105	.064	.021
12	.107	.066	.019	.055	.032	.007	.008	.004	.000
16	.131	.060	.011	.058	.030	.006	.022	.008	.001
19	.134	.035	.000	.170	.052	.008	.692	.053	.013

higher than expected, or even exceeded .50 (Table 5; see, for example, for $X_+ \leq 17$ and $X_+ \geq 75$ the Type I error rates at significance level of .10).

Tables 3 through 5 also show that for large spread of the δ s, the differences between nominal Type I error rates and simulated Type I error rates increased much faster in the tails of the X_+ distribution, compared with the simulated Type I error rates for tests with small spread of δ s. See, for example, the simulated Type I error rates at $X_+ = 25$, for moderate item discrimination (Table 5). For small spread of item difficulties, these Type I error rates were 0.193, 0.103, and 0.018 at significance levels of .10, .05, and .01, respectively. For large spread of item difficulties, the corresponding Type I error rates were much higher: 0.449, 0.322, and 0.135, respectively.

For the RM, Tables 6 through 8 show the simulated Type I error rates at three significance levels. Compared with the 3PLM, we found similar trends for the Type I error rates as the item discrimination increased. More specifically, increasing the item discrimination yielded Type I error rates that were smaller than expected for X_+ in the middle of the X_+ distribution and larger in the tails of the X_+ distribution. For example, in Table 8 for $J = 80$, small spread of the δ s, and a significance level of .10, the Type I error rates for $25 \leq X_+ \leq 65$ varied from 0.112 to 0.157 (weak discrimination), 0.091 to 0.200 (moderate discrimination), and 0.003 to 0.080 (strong discrimination). In addition, for $X_+ \leq 15$ and $X_+ \geq 70$, the Type I error rates varied from 0.188 to 1.000 (moderate discrimination) and 0.084 to 1.000 (strong discrimination). As with the 3PLM, no results were obtained for the highly improbable very low and very high X_+ values under weak item discrimination.

Again, these results could well be explained by a shift of the simulated $ZU3$ distribution relative to the expected $ZU3$ distribution.

Results of the simulations using the monotone homogeneity model. Tables 9 through 11 show the simulated Type I error rates for the simulations under the MHM. In general, similar results were found for the MHM as for the 3PLM and the RM. More specifically, the results for the condition

Table 4
Simulated Type I Error Rates (Three-Parameter Logistic Model)
at Three Significance Levels (Sign. Lev.), for $J = 40$, Three
Levels of Item Discrimination, and Two Levels of Spread of
Item Difficulties (1,000 Observations at Each Score Level)

X_+	Item Discrimination								
	Weak			Moderate			Strong		
	Sign. Lev.			Sign. Lev.			Sign. Lev.		
	.10	.05	.01	.10	.05	.01	.10	.05	.01
Small Spread of Item Difficulties									
1	—	—	—	.829	.418	.008	1.000	1.000	.680
4	—	—	—	.482	.273	.041	.886	.713	.302
8	.122	.058	.013	.228	.125	.030	.455	.314	.115
12	.108	.050	.008	.129	.066	.014	.136	.075	.018
16	.104	.063	.014	.074	.038	.005	.044	.020	.002
20	.096	.049	.010	.057	.021	.004	.024	.006	.000
24	.088	.045	.009	.052	.020	.005	.008	.002	.002
28	.118	.065	.011	.071	.030	.009	.008	.003	.000
32	.148	.060	.008	.093	.046	.006	.033	.016	.000
36	.149	.045	.000	.196	.069	.002	.135	.036	.004
39	—	—	—	.293	.038	.000	.805	.199	.007
Large Spread of Item Difficulties									
1	—	—	—	1.000	1.000	1.000	1.000	1.000	1.000
4	—	—	—	.996	.954	.730	1.000	1.000	1.000
8	.339	.206	.045	.683	.523	.275	.966	.899	.698
12	.182	.094	.022	.304	.202	.080	.459	.328	.148
16	.094	.047	.015	.121	.071	.017	.084	.037	.013
20	.080	.038	.011	.045	.016	.004	.010	.003	.000
24	.078	.040	.011	.035	.015	.000	.007	.003	.000
28	.109	.059	.017	.049	.022	.004	.005	.003	.000
32	.156	.080	.017	.124	.062	.008	.024	.005	.000
36	.380	.182	.023	.510	.270	.041	.843	.421	.047
39	—	—	—	1.000	1.000	.178	1.000	1.000	1.000

Note. Lines indicate that the Type I error rate was not obtained because these levels of X_+ had approximately zero probability.

“weak discrimination,” as defined under the MHM, were comparable with the results obtained in the condition of “moderate discrimination,” as defined under the 3PLM and the RM. Furthermore, as can be seen in Tables 9 through 11, the Type I error rates somewhat increased in the middle range of X_+ .

Discussion

A sampling distribution of $U3$ is needed for identifying examinees with significantly deviant item score patterns when there is no a priori knowledge of the percentage of misfitting patterns. This study explored the applicability of the theoretical distribution of $U3$ for classifying nonfitting item score patterns. We investigated the robustness of the assumption that the standardized version of $U3$, $ZU3$, follows a standard normal distribution. In particular, we investigated whether standard normal deviates for $ZU3$ are suitable for identifying misfitting item score patterns at a nominal significance level.

It was shown that as the item discrimination increased, the simulated $ZU3$ distributions differed more from the standard normal distribution and, consequently, the Type I error rates were either too high or too low to be used in practice. Differences between the theoretical and simulated

Table 5
Simulated Type I Error Rates (Three-Parameter Logistic Model)
at Three Significance Levels (Sign. Lev.), for $J = 80$, Three
Different Levels of Discrimination, and Two Levels of Spread
of Item Difficulties (1,000 Observations at Each Score Level)

X_+	Item Discrimination								
	Weak			Moderate			Strong		
	Sign. Lev.			Sign. Lev.			Sign. Lev.		
	.10	.05	.01	.10	.05	.01	.10	.05	.01
Small Spread of Item Difficulties									
5	—	—	—	—	—	—	1.000	1.000	.970
10	—	—	—	.755	.538	.158	.975	.922	.665
15	—	—	—	.531	.353	.112	.687	.535	.250
20	.243	.126	.025	.322	.192	.052	.286	.173	.059
25	.150	.083	.010	.193	.103	.018	.085	.045	.006
30	.123	.066	.014	.108	.066	.020	.017	.008	.002
35	.113	.055	.013	.071	.040	.010	.007	.003	.001
40	.116	.067	.017	.044	.018	.005	.002	.000	.000
45	.082	.033	.009	.040	.018	.002	.000	.000	.000
50	.080	.033	.007	.042	.012	.002	.001	.000	.000
55	.094	.049	.008	.039	.020	.002	.001	.000	.000
60	.075	.029	.004	.055	.016	.004	.009	.004	.000
65	.082	.030	.002	.099	.037	.005	.031	.010	.000
70	.110	.032	.002	.160	.061	.009	.131	.051	.002
75	—	—	—	.364	.137	.005	.672	.343	.032
79	—	—	—	.912	.288	.000	1.000	1.000	.295
Large Spread of Item Difficulties									
5	—	—	—	—	—	—	1.000	1.000	1.000
10	—	—	—	1.000	.999	.983	1.000	1.000	1.000
15	—	—	—	.976	.926	.763	1.000	1.000	.977
20	.560	.410	.146	.770	.662	.381	.931	.861	.667
25	.365	.224	.081	.449	.322	.135	.471	.351	.159
30	.213	.127	.034	.220	.131	.026	.111	.064	.017
35	.124	.070	.021	.101	.049	.014	.017	.007	.001
40	.107	.051	.007	.038	.018	.003	.004	.001	.000
45	.074	.037	.009	.030	.017	.003	.000	.000	.000
50	.066	.031	.005	.024	.008	.001	.000	.000	.000
55	.070	.038	.006	.033	.014	.002	.000	.000	.000
60	.111	.050	.009	.054	.020	.004	.008	.002	.000
65	.187	.092	.006	.166	.075	.010	.149	.053	.001
70	.326	.152	.016	.555	.314	.062	.967	.752	.189
75	—	—	—	.997	.927	.400	1.000	1.000	1.000
79	—	—	—	—	—	—	1.000	1.000	1.000

Note. Lines indicate that the Type I error rate was not obtained because these levels of X_+ had approximately zero probability.

distributions may be due to the inadequacy of the regression formulas (Equations 3 and 4) to obtain theoretical expressions for the mean and the standard deviation of the conditional sampling distribution of U_3 . These regression formulas were used to predict the conditional distribution of $W(\mathbf{X})$ given X_+ and relied on the assumption that X_+ and $W(\mathbf{X})$ follow a bivariate normal distribution. However, as the item discrimination increased, the unconditional X_+ distribution deviated increasingly from a normal distribution (see, for example, Lord & Novick, 1968, p. 388).

Table 6
Simulated Type I Error Rates (Rasch Model) at Three Significance Levels (Sign. Lev.), for $J = 20$, Three Different Levels of Discrimination, and Two Levels of Spread of Item Difficulties (1,000 Observations at Each Score Level)

X_+	Item Discrimination								
	Weak			Moderate			Strong		
	Sign. Lev.			Sign. Lev.			Sign. Lev.		
	.10	.05	.01	.10	.05	.01	.10	.05	.01
Small Spread of Item Difficulties									
1	.064	.000	.000	.181	.053	.000	.302	.113	.009
4	.093	.045	.006	.105	.047	.005	.053	.022	.003
8	.084	.038	.011	.060	.029	.009	.015	.005	.001
12	.107	.054	.009	.070	.040	.012	.009	.004	.001
16	.088	.038	.002	.104	.043	.006	.047	.021	.006
19	.038	.000	.000	.157	.056	.000	.228	.115	.007
Large Spread of Item Difficulties									
1	.262	.099	.000	1.000	.662	.125	1.000	1.000	1.000
4	.122	.053	.005	.158	.082	.012	.156	.071	.008
8	.107	.055	.010	.048	.028	.006	.008	.001	.001
12	.085	.043	.007	.055	.028	.011	.006	.002	.000
16	.124	.064	.009	.166	.073	.018	.147	.073	.009
19	.208	.097	.000	1.000	.443	.099	1.000	1.000	1.000

Table 7
Simulated Type I Error Rates (Rasch Model) at Three Significance Levels (Sign. Lev.), for $J = 40$, Three Different Levels of Discrimination, and Two Levels of Spread of Item Difficulties (1,000 Observations at Each Score Level)

X_+	Item Discrimination								
	Weak			Moderate			Strong		
	Sign. Lev.			Sign. Lev.			Sign. Lev.		
	.10	.05	.01	.10	.05	.01	.10	.05	.01
Small Spread of Item Difficulties									
1	—	—	—	.480	.185	.000	1.000	1.000	.163
4	.105	.026	.000	.249	.092	.008	.376	.164	.013
8	.105	.050	.006	.116	.048	.014	.058	.026	.001
12	.105	.048	.007	.075	.031	.007	.017	.007	.002
16	.097	.046	.015	.064	.030	.004	.007	.004	.001
20	.073	.044	.007	.063	.028	.005	.004	.003	.001
24	.093	.046	.006	.057	.027	.003	.004	.001	.000
28	.101	.040	.005	.066	.026	.005	.010	.003	.000
32	.107	.043	.011	.127	.066	.015	.044	.023	.001
36	.129	.048	.003	.230	.107	.011	.309	.127	.016
39	—	—	—	.477	.198	.000	1.000	1.000	.120
Large Spread of Item Difficulties									
1	—	—	—	1.000	1.000	1.000	1.000	1.000	1.000
4	.331	.163	.028	.839	.572	.172	1.000	1.000	1.000
8	.180	.090	.020	.223	.118	.017	.355	.166	.015
12	.133	.066	.008	.086	.039	.008	.022	.008	.001
16	.097	.051	.012	.039	.023	.004	.000	.000	.000
20	.095	.044	.009	.034	.014	.001	.005	.003	.000
24	.083	.037	.005	.056	.026	.004	.002	.002	.000
28	.118	.063	.012	.089	.035	.009	.015	.003	.001
32	.164	.074	.015	.300	.159	.033	.327	.140	.025
36	.344	.174	.026	.918	.721	.252	1.000	1.000	.966
39	—	—	—	1.000	1.000	1.000	1.000	1.000	1.000

Note. Lines indicate that the Type I error rate was not obtained because these levels of X_+ had approximately zero probability.

Table 8
Simulated Type I Error Rates (Rasch Model) at Three Significance Levels (Sign. Lev.), for $J = 80$, Three Different Levels of Discrimination, and Two Levels of Spread of Item Difficulties (1,000 Observations at Each Score Level)

X_+	Item Discrimination								
	Weak			Moderate			Strong		
	Sign. Lev.			Sign. Lev.			Sign. Lev.		
	.10	.05	.01	.10	.05	.01	.10	.05	.01
Small Spread of Item Difficulties									
5	—	—	—	—	—	—	1.000	.943	.320
10	—	—	—	.344	.189	.032	.433	.214	.034
15	—	—	—	.188	.094	.013	.084	.037	.003
20	.136	.062	.011	.106	.039	.006	.017	.004	.000
25	.112	.048	.004	.091	.041	.006	.003	.000	.000
30	.079	.038	.007	.055	.019	.003	.000	.000	.000
35	.082	.041	.002	.049	.017	.001	.002	.000	.000
40	.089	.044	.008	.045	.018	.005	.002	.000	.000
45	.080	.035	.011	.054	.025	.003	.002	.001	.000
50	.093	.052	.011	.054	.023	.004	.000	.000	.000
55	.119	.050	.005	.079	.037	.003	.003	.000	.000
60	.126	.064	.009	.100	.037	.007	.024	.006	.000
65	.157	.053	.004	.200	.102	.013	.080	.028	.001
70	.166	.077	.005	.337	.165	.015	.409	.187	.029
75	—	—	—	.718	.414	.054	1.000	.920	.309
79	—	—	—	1.000	1.000	.163	1.000	1.000	1.000
Large Spread of Item Difficulties									
5	—	—	—	—	—	—	1.000	1.000	1.000
10	—	—	—	.966	.855	.450	1.000	1.000	1.000
15	—	—	—	.530	.346	.103	.917	.666	.212
20	.185	.093	.019	.186	.094	.019	.122	.043	.007
25	.137	.066	.017	.077	.035	.008	.006	.003	.000
30	.082	.038	.007	.046	.019	.008	.000	.000	.000
35	.077	.041	.008	.027	.011	.001	.000	.000	.000
40	.095	.049	.012	.024	.009	.000	.000	.000	.000
45	.099	.045	.008	.026	.014	.001	.000	.000	.000
50	.089	.044	.009	.044	.022	.002	.000	.000	.000
55	.140	.066	.017	.076	.041	.008	.015	.004	.000
60	.202	.115	.020	.224	.116	.028	.141	.054	.006
65	.288	.160	.021	.578	.369	.125	.932	.736	.245
70	—	—	—	.976	.881	.485	1.000	1.000	1.000
75	—	—	—	1.000	1.000	1.000	1.000	1.000	1.000
79	—	—	—	—	—	—	1.000	1.000	1.000

Note. Lines indicate that the Type I error rate was not obtained because these levels of X_+ had approximately zero probability.

Table 9
Simulated Type I Error Rates (Monotone Homogeneity Model) at Three Significance Levels (Sign. Lev.), for $J = 20$, Three Levels of Item Discrimination, and Two Levels of Spread of Item Difficulties (1,000 Observations at Each Score Level)

X_+	Item Discrimination								
	Weak			Moderate			Strong		
	Sign. Lev.			Sign. Lev.			Sign. Lev.		
	.10	.05	.01	.10	.05	.01	.10	.05	.01
Small Spread of Item Difficulties									
1	.120	.016	.000	.119	.000	.000	.012	.000	.000
4	.038	.011	.004	.026	.006	.001	.016	.007	.000
8	.041	.020	.003	.038	.017	.002	.023	.011	.001
12	.084	.035	.003	.068	.027	.007	.046	.020	.004
16	.060	.023	.003	.050	.018	.000	.040	.013	.000
19	.123	.003	.000	.054	.000	.000	.000	.000	.000
Large Spread of Item Difficulties									
1	.761	.352	.034	.568	.191	.006	.243	.059	.000
4	.083	.034	.004	.037	.020	.001	.027	.012	.000
8	.035	.017	.006	.020	.007	.003	.006	.004	.000
12	.024	.007	.001	.011	.004	.002	.012	.007	.001
16	.098	.034	.005	.064	.015	.000	.038	.015	.000
19	.417	.245	.001	.314	.146	.001	.165	.047	.000

Table 10
Simulated Type I Error Rates (Monotone Homogeneity Model) at Three Significance Levels (Sign. Lev.), for $J = 40$, Three Levels of Item Discrimination, and Two Levels of Spread of Item Difficulties (1,000 Observations at Each Score Level)

X_+	Item Discrimination								
	Weak			Moderate			Strong		
	Sign. Lev.			Sign. Lev.			Sign. Lev.		
	.10	.05	.01	.10	.05	.01	.10	.05	.01
Small Spread of Item Difficulties									
1	.145	.073	.000	.095	.033	.000	.080	.000	.000
4	.099	.031	.002	.043	.015	.001	.034	.013	.002
8	.018	.007	.000	.015	.003	.001	.010	.003	.000
12	.066	.037	.008	.041	.022	.004	.025	.013	.002
16	.094	.052	.013	.078	.036	.012	.062	.026	.002
20	.102	.048	.010	.067	.032	.004	.065	.030	.002
24	.056	.017	.004	.057	.023	.002	.036	.016	.004
28	.043	.022	.003	.041	.020	.000	.035	.009	.000
32	.034	.010	.001	.020	.004	.000	.008	.001	.000
36	.025	.008	.000	.015	.005	.000	.014	.005	.000
39	.114	.010	.000	.049	.000	.000	.017	.000	.000
Large Spread of Item Difficulties									
1	1.000	1.000	.654	1.000	.795	.267	.718	.350	.002
4	.708	.513	.201	.577	.381	.096	.300	.145	.017
8	.141	.081	.011	.058	.025	.001	.024	.006	.001
12	.027	.008	.000	.015	.003	.000	.000	.000	.000
16	.008	.002	.000	.004	.000	.000	.004	.001	.000
20	.011	.002	.000	.008	.001	.000	.000	.000	.000
24	.026	.010	.003	.016	.007	.001	.003	.001	.000
28	.058	.023	.003	.015	.007	.000	.011	.003	.000
32	.213	.109	.017	.112	.051	.007	.020	.006	.001
36	.774	.541	.183	.593	.371	.105	.316	.155	.029
39	1.000	1.000	.750	1.000	1.000	.393	.882	.463	.023

Table 11
Simulated Type I Error Rates (Monotone Homogeneity Model) at Three Significance Levels (Sign. Lev.), for $J = 80$, Three Levels of Item Discrimination, and Two Levels of Spread of Item Difficulties (1,000 Observations at Each Score Level)

X_+	Item Discrimination								
	Weak			Moderate			Strong		
	Sign. Lev.			Sign. Lev.			Sign. Lev.		
	.10	.05	.01	.10	.05	.01	.10	.05	.01
Small Spread of Item Difficulties									
5	.493	.216	.017	.333	.122	.006	.111	.024	.001
10	.103	.037	.002	.052	.016	.002	.016	.004	.000
15	.014	.005	.000	.005	.000	.000	.004	.000	.000
20	.014	.011	.002	.009	.002	.000	.003	.001	.000
25	.041	.018	.002	.030	.013	.001	.020	.007	.000
30	.098	.047	.005	.043	.018	.005	.030	.016	.000
35	.103	.054	.014	.082	.035	.012	.037	.014	.003
40	.104	.048	.007	.092	.040	.009	.075	.029	.006
45	.078	.041	.011	.066	.035	.007	.040	.016	.006
50	.058	.022	.004	.059	.022	.001	.035	.018	.005
55	.043	.020	.001	.048	.020	.003	.025	.011	.000
60	.030	.009	.001	.029	.005	.000	.021	.004	.000
65	.024	.005	.000	.011	.003	.000	.008	.001	.000
70	.031	.005	.000	.017	.002	.000	.005	.002	.000
75	.218	.066	.003	.134	.034	.000	.061	.009	.000
79	.601	.132	.000	.398	.015	.000	.119	.001	.000
Large Spread of Item Difficulties									
5	1.000	1.000	.962	.999	.982	.735	.853	.641	.161
10	.855	.647	.263	.616	.390	.108	.172	.068	.006
15	.297	.139	.018	.101	.039	.007	.009	.003	.000
20	.052	.022	.001	.013	.003	.000	.000	.000	.000
25	.014	.005	.000	.002	.000	.000	.000	.000	.000
30	.007	.001	.000	.002	.002	.000	.001	.000	.000
35	.007	.003	.000	.005	.002	.000	.000	.000	.000
40	.010	.004	.001	.004	.002	.000	.000	.000	.000
45	.013	.006	.000	.008	.004	.000	.000	.000	.000
50	.016	.005	.000	.001	.000	.000	.002	.000	.000
55	.026	.009	.002	.006	.004	.000	.000	.000	.000
60	.053	.021	.000	.015	.004	.000	.004	.000	.000
65	.141	.061	.005	.052	.020	.003	.007	.002	.001
70	.626	.409	.127	.385	.206	.039	.126	.050	.009
75	.999	.979	.762	.984	.912	.499	.831	.598	.108
79	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.345

Consequently, the assumption of a bivariate normally distributed X_+ and $W(\mathbf{X})$ was violated and the conditional distribution of $W(\mathbf{X})$ could not be accurately estimated.

The conclusion is that the theoretical sampling distribution of U_3 should not be used for testing hypotheses about item score patterns. However, U_3 can be used for ordering item score patterns according to their likelihood (van der Flier, 1980). This means that if one wishes to select a percentage of the most improbable patterns, U_3 provides a useful descriptive statistic. In fact, Meijer et al. (1994) demonstrated that an increasing item discrimination yielded higher detection rates of nonfitting item score patterns, in particular for long tests (at least 33 items).

Finally, recent studies have compared the theoretical and simulated distributions of person-fit statistics in the context of parametric IRT (Reise, 1995; Nering, 1997; van Krimpen-Stoop & Meijer, 1999; Snijders, 2001). The results of these studies are in some way comparable with the results of this study: It was found that in the middle of the θ range, simulated and nominal Type I error rates were similar, but that for extreme θ larger differences existed.

References

- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 396-479). Reading, MA: Addison-Wesley.
- 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.
- Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction* (pp. 60-90). Princeton, NJ: Princeton University Press.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.
- Harnisch, D. L., & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement*, 18, 133-146.
- Hemker, B. T., Sijtsma, K., & Molenaar, I. W. (1995). Selection of unidimensional scales from a multidimensional item bank in the polytomous Mokken IRT model. *Applied Psychological Measurement*, 19, 337-352.
- 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.
- Junker, B. W. (1993). Conditional association, essential independence and monotone unidimensional item response models. *The Annals of Statistics*, 21, 1359-1378.
- Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple choice test scores. *Journal of Educational Statistics*, 4, 269-290.
- Lindgren, B. W. (1993). *Statistical theory*. New York: Chapman & Hall.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Meijer, R. R. (1994). *Nonparametric person fit analysis*. Unpublished doctoral thesis, Vrije Universiteit, Amsterdam.
- 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. (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.
- 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*. New York/Berlin: De Gruyter.
- Mokken, R. J., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement*, 6, 417-430.
- Mokken, R. J., Lewis, C., & Sijtsma, K. (1986). Rejoinder to "The Mokken scale: A critical discussion." *Applied Psychological Measurement*, 10, 279-285.

- Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person-fit indices. *Psychometrika*, 55, 75-106.
- Nering, M. L. (1997). The distribution of indexes of person fit within the computerized adaptive testing environment. *Applied Psychological Measurement*, 21, 115-127.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.
- Reise, S. P. (1995). Scoring method and the detection of person misfit in a personality assessment context. *Applied Psychological Measurement*, 19, 213-229.
- Roskam, E. E., van den Wollenberg, A. L., & Jansen, P. G. W. (1986). The Mokken scale: A critical discussion. *Applied Psychological Measurement*, 10, 265-277.
- 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.
- Sijsma, K. (1998). Methodology review: Nonparametric IRT approaches to the analysis of dichotomous item scores. *Applied Psychological Measurement*, 22, 3-31.
- Sijsma, K., & Meijer, R. R. (1992). A method for investigating the intersection of item response functions in Mokken's nonparametric IRT model. *Applied Psychological Measurement*, 16, 149-157.
- Snijders, T. (2001). Asymptotic null distribution of person-fit statistics with estimated person-parameter. *Psychometrika*, 66, 331-342.
- 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.
- van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (1999). The null distribution of person-fit statistics for conventional and adaptive tests. *Applied Psychological Measurement*, 23, 327-345.

Authors' Addresses

Send requests for reprints or further information to Wilco H. M. Emons, Department of Methodology and Statistics, FSW, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, the Netherlands; Rob R. Meijer, TO/OMD, P.O. Box 217, 7500 AE Enschede, the Netherlands; Klaas Sijsma, Department of Methodology and Statistics, FSW, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, the Netherlands. E-mail: w.h.m.emons@kub.nl; meijer@edte.utwente.nl; k.sijtsma@kub.nl.