DOCUMENT RESUME

ED 421 548 TM 028 874

AUTHOR	Meijer, Rob R.; van Krimpen-Stoop, Edith M. L. A.
TITLE	Simulating the Null Distribution of Person-Fit Statistics for Conventional and Adaptive Tests. Research Report 98-02.
INSTITUTION	Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.
PUB DATE	1998-00-00
NOTE	40p.
AVAILABLE FROM	Faculty of Educational Science and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
PUB TYPE	Reports - Evaluative (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Ability; *Adaptive Testing; Foreign Countries; Item Response Theory; Responses; Scores; *Simulation; *Statistical Distributions; *Test Items
IDENTIFIERS	*Person Fit Measures; Power (Statistics); Two Parameter Model

ABSTRACT

Several person-fit statistics have been proposed to detect item score patterns that do not fit an item response theory model. To classify response patterns as not fitting a model, a distribution of a person-fit statistic is needed. The null distributions of several fit statistics have been investigated using conventionally administered tests, but less is known about the distribution of fit statistics for computerized adaptive testing (CAT). A three-part simulation to study this distribution is described. First the theoretical distribution of the often used l(z) statistic across theta levels in a conventional testing and in CAT testing was studied, where theta and estimated theta were used to calculate l(z). Also, the distribution of a statistic l*(z), that is corrected for the error in theta, proposed by T. Snijders (1998) was studied in both testing environments. Simulating the distribution of l(z) for the two-parameter logistic model for conventional tests was studied. Two procedures for simulating the distribution of l(z) and l*(z) in a CAT were examined: (1) item scores were simulated with a fixed set of administered items; and (2) item scores were generated according to a stochastic design, where the choice of the administered item i + 1 depended on responses to previously administered items. The third study was a power study conducted to compare detection rates of l*(z) with l(z) for conventional tests. Results indicate that the distribution of l(z) differed from the theoretical distribution in conventional and CAT environments. In a conventional testing situation, the distribution of l(z) was in accord with the theoretical distribution, but for the CAT the distribution differed from the theoretical distribution. In the context of conventional testing, simulating the sampling distribution of l(z) for every examinee, based on theta, resulted in an appropriate approximation of the distribution. However, for the CAT environment, simulating the sampling distributions of both l(z) and l*(z) was problematic. Two appendixes show the derivation of the l*(z) statistic and discuss modeling local dependence. (Contains 6 tables, 3 figures, and 24 references.) (Author/SLD)

Simulating the Null Distribution of Person-Fit Statistics for Conventional and Adaptive Tests

Research Report 98-02

Rob R. Meijer
Edith M.L.A. van Krimpen-Stoop

faculty of
EDUCATIONAL SCIENCE AND TECHNOLOGY

University of Twente

Department of
Educational Measurement and Data Analysis

Simulating the Null Distribution of Person-Fit Statistics for Conventional and Adaptive Tests

Rob R. Meijer

Edith M. L. A. van Krimpen-Stoop

Abstract

Several person-fit statistics have been proposed to detect item score patterns that do not fit an item response theory model. To classify response patterns as not fitting a model a distribution of a person-fit statistic is needed. Recently, the null distributions of several fit statistics have been investigated using conventional administered tests. For computerized adaptive testing (CAT), however, less is known about the distribution of fit statistics. In this study a three part simulation study was conducted. First, the theoretical distribution of the often used l_z-statistic across θ-levels in a conventional testing and CAT environment was investigated, where θ and $\widehat{\theta}$ were used to calculate l_z. Also, the distribution of a statistic l_z^*, that is corrected for the error in $\hat{\theta}$, proposed by Snijders (1998), was investigated in a conventional testing and CAT environment. Second, simulating the distribution of l_z for the 2PLM for conventional administered tests was investigated. Two procedures for simulating the distribution of l_z and l_z^* in a CAT were examined: (1) item scores were simulated with a fixed set of administered items, and (2) item scores were generated according to a stochastic design, where the choice of the administered item $i + 1$ depended on the responses to previous administered items. Third, a power study was conducted to compare the detection rates of l_z^* with l_z for conventional tests. Results indicated that the distribution of l_z differed from the theoretical distribution in a conventional and CAT environment. In a conventional testing situation, the distribution of l_z^* was in concordance with the theoretical distribution. However, for a CAT the distribution differed from the theoretical distribution. In the context of conventional testing, simulating the sampling distribution of l_z for every examinee, based on $\hat{\theta}$, resulted in an appropriate approximation of the distribution. However, in a CAT environment, simulating the sampling distributions of both l_z and l_z^* was problematic.

Simulating the Null Distribution of Person-Fit Statistics for Conventional and Adaptive Tests

Item responses that do not fit the assumed item response theory (IRT) model may cause the latent trait value θ, to be inaccurately estimated. Possible interpretations for nonfitting test behavior include test anxiety, guessing, cheating on achievement tests, or response distortion as a result of faking the answers on personality inventories (Zickar & Drasgow, 1996). Person-fit statistics have been proposed to detect nonfitting score patterns (e.g., Drasgow & Levine, 1986; Meijer, 1994; Tatsuoka, 1984), and the effectiveness of these statistics to detect nonfitting response vectors has been investigated (e.g., Drasgow, Levine, & McLaughlin, 1987, 1991). However, most person-fit studies concentrated on conventionally administered tests, or paper-and-pencil (P&P) tests. With the increasing use of computerized adaptive tests (CAT), additional research is needed with respect to the application of person-fit statistics using these types of tests.

A few studies investigated the usefulness of person-fit analysis in CAT. Candell (1988; cited in Drasgow, Levine, and Zickar, 1996) used optimal person-fit statistics in which the likelihood of a normal responding person was compared with the likelihood under an aberrant model to study the ability of a likelihood ratio test to identify simulated aberrant examinees. Although this approach is interesting because it has maximum power against a specified alternative, the drawback is that for other types of aberrant responding the power is low.

Nering (1997) examined the distribution of two fit statistics, l_z (Drasgow, Levine, & Williams, 1985), a standardized version of the log-likelihood statistic l_0 proposed by Levine and Rubin (1979), and $ECI4_z$ (Tatsuoka, 1984), in a CAT environment. Nering found that the empirical distribution was dramatically different from the theoretical distribution. As a result, Nering concluded that 'when attempting to classify a response vector as model divergent (...) cutscores may have to be based on such factors as item pool size, item pool discrimination, and so forth'. An alternative to using a critical value derived from a theoretical distribution is to simulate for each examinee a distribution of a person-fit statistic based on the characteristics of the item bank and the estimated latent trait value of θ, denoted as $\widehat{\theta}$. Using the simulated distribution it can be determined how likely a response vector is under the IRT model.

Snijders (1998) proposed to use an alternative standardization of the l_0-statistic when $\hat{\theta}$ was replaced by θ. This statistic is denoted here as l_z^*. In a small simulation study Snijders (1998) investigated the distribution of l_z^* in a conventional testing environment and found that the empirical distribution was close to the theoretical distribution.

The purpose of the present study was to extend the Nering study and the Snijders study by examining the distribution of l_z and l_z^* in a conventional testing and CAT environment and to investigate two different ways to simulate the distribution of l_0, l_z, and l_z^*. Besides, the detection rate of l_z and l_z^* to detect nonfitting score patterns for conventional tests was investigated.

Person-Fit Analysis

In person-fit analysis, several fit statistics have been used in the context of the one−, two-, and three-parameter logistic model (1-, 2-, 3PLM) (Hambleton & Swaminatan, 1985, pp. 35-48). In this study we use the 2PLM because it is less restrictive with respect to empirical data than the one-parameter logistic model and it does not have the estimation problems of the guessing parameter in the three parameter logistic model (e.g., Baker, 1992, pp.109-112). The 2PLM has shown to have a reasonable fit to several achievement and personality data (e.g., Reise & Waller, 1990; Zickar & Drasgow, 1996).

Let X_i be the binary $(0, 1)$ response to item i, where 1 denotes a correct or keyed response, and 0 denotes an incorrect or not keyed response. Further, let a_i denote the item discrimination parameter and b_i the item difficulty parameter, then the probability of correctly answering an item according to the 2PLM can be written as

$$P_i(\theta) = \frac{\exp[a_i(\theta - b_i)]}{1 + \exp[a_i(\theta - b_i)]}. \tag{1}$$

Levine and Rubin (1979) proposed the log-likelihood statistic, denoted as l_0, as a measure of departure from the logistic IRT models. l_0 can be written as

$$l_0 = \ln\left[\prod_{i=1}^{n} P_i\left(\hat{\theta}\right)^{X_i}\left(1 - P_i\left(\hat{\theta}\right)\right)^{1-X_i}\right]. \tag{2}$$

Because l_0 is confounded with $\widehat{\theta}$, Drasgow, Levine and Williams (1985) proposed to use the standardized version of l_0, denoted as l_z. This statistic equals

$$l_z = \frac{l_0 - E(l_0)}{[\mathrm{var}(l_0)]^{\frac{1}{2}}}, \tag{3}$$

where $E(l_0)$ and $\mathrm{var}(l_0)$ denote the expectation and variance of l_0, respectively. These

quantities are given by

$$E(l_0) = \sum_{i=1}^{n} \left\{ P_i\left(\hat{\theta}\right) \ln\left[P_i\left(\hat{\theta}\right)\right] + \left(1 - P_i\left(\hat{\theta}\right)\right) \ln\left[1 - P_i\left(\hat{\theta}\right)\right] \right\}, \quad (4)$$

and

$$\text{var}(l_0) = \sum_{i=1}^{n} P_i\left(\hat{\theta}\right)\left(1 - P_i\left(\hat{\theta}\right)\right)\left[\ln \frac{P_i(\hat{\theta})}{1-P_i(\hat{\theta})}\right]^2. \quad (5)$$

For classifying a response pattern as aberrant, an important tool is the probability of exceedance or significance probability. Because large negative values of l_z indicate aberrance, the significance probabilities in the left tail of the distribution are of interest. Let t be the observed value of the person-fit statistic T. Then, the significance probability is defined as the probability under the sampling distribution that the value of the test statistic is smaller than the observed value of the statistic: $p^* = P(T \leq t)$. The value of a statistic with $p^* = \alpha$ will be denoted as the critical value at significance level α. For example, for a standard normally distributed statistic the critical value at level $\alpha = 0.05$ is -1.65.

Drasgow et al. (1985) purported that, in the context of conventional testing or paper-and-pencil (P&P) testing, l_z was distributed standard normal for long tests (tests longer than say 80 items). However, several studies (e.g., Molenaar & Hoijtink, 1990; Meijer & Nering, 1997) showed that l_z was not standard normally distributed for tests of realistic length ($20 - 60$ items). It was found that the distribution of l_z was negatively skewed and that the normal approximation was inaccurate, especially in the tails of the distribution. As an alternative, Molenaar and Hoijtink (1990) proposed for the Rasch model three approximations to the distribution of l_0, conditional on the total score: using (1) complete enumeration, (2) Monte Carlo simulation and (3) a chi-square distribution, where the mean, standard deviation, and skewness of l_0 were taken into account. Complete enumeration is suitable for very short tests. For tests of moderate length, a chi-square distribution was proposed for l_0, conditional on the total score. For very long tests, an accurate calculation of the moments, needed for the chi-square approximation, is difficult and as an alternative, Monte Carlo simulation was applied. In the Rasch model the total score is a sufficient statistic for θ; for the 2PLM this is not the case, that is, the distribution of a person-fit statistic conditional on the total score is dependent on θ. As an alternative $\widehat{\theta}$ can be used in the case of the 2PLM or 3PLM. However, care should be taken in doing this, because as Molenaar and Hoijtink (1990) noticed the statistical results on the distribution of a person-fit

statistic may change when substituting $\widehat{\theta}$ for θ.

Snijders (1998; see also Molenaar and Hoijtink, 1990) showed that, when the true person parameter is replaced by an estimate, the variance of the person-fit statistic decreased. When this decreased variance is not taken into account, this will lead to a conservative classification of nonfitting response patterns. Snijders derived the asymptotic distribution for several person-fit statistics, in which θ was replaced by $\hat{\theta}$. He showed that the asymptotic distribution of

$$l_z^* = \frac{l_0 - E(l_0) + c_n\left(\hat{\theta}\right) r_0\left(\hat{\theta}\right)}{\sqrt{n}\tau_n\left(\hat{\theta}\right)} \tag{6}$$

is standard normal, where, for the 2PLM and for the weighted maximum likelihood estimator (Warm, 1989)

$$c_n\left(\hat{\theta}\right) = \frac{\sum_{i=1}^{n} a_i\left(\hat{\theta} - b_i\right) P_i'\left(\hat{\theta}\right)}{\sum_{i=1}^{n} a_i P_i'\left(\hat{\theta}\right)} \tag{7}$$

$$r_0\left(\hat{\theta}\right) = \frac{\sum_{i=1}^{n} a_i^3 P_i\left(\hat{\theta}\right)\left[1 - P_i\left(\hat{\theta}\right)\right]\left[1 - 2P_i\left(\hat{\theta}\right)\right]}{2\sum_{i=1}^{n} a_i^2 P_i\left(\hat{\theta}\right)\left[1 - P_i\left(\hat{\theta}\right)\right]} \tag{8}$$

$$\tau_n^2\left(\hat{\theta}\right) = \frac{1}{n}\sum_{i=1}^{n}\left[a_i\left(\hat{\theta} - b_i\right) - a_i c_n\left(\hat{\theta}\right)\right] P_i\left(\hat{\theta}\right)\left[1 - P_i\left(\hat{\theta}\right)\right] \tag{9}$$

where $P_i'(\theta) = \partial P_i(\theta)/\partial\theta$. See Appendix A for more details. Snijders performed a simulation study for relatively small conventional tests of 8 and 15 items, fitting the 2PLM, and using maximum likelihood estimation. The results showed that the approximation is satisfactory for $\alpha = 0.05$ and $\alpha = 0.10$, but that it was too liberal for smaller values of α.

Person Fit in CAT

Nering (1996, 1997) examined the distribution of l_z within CAT by evaluating the first four moments (mean, standard deviation, skewness, and kurtosis) of the distribution. His results were in concordance with the results using conventional tests: the exact distribution of l_z was different from the standard normal distribution. An important finding was that the normal approximation in the tails of the distribution was inaccurate: using a critical value of -1.65

resulted in a conservative classification of aberrant response patterns: $p^* = P(l_z \leq -1.65) \ll 0.05$. On the basis of these results it can be concluded that the standard normal distribution is not useful to obtain a critical value in CAT. A possible solution for determining significance probabilities in CAT may be to approximate the distribution of a statistic by using simulation methods.

Simulating Distributions of Person-Fit Statistics in CAT

The significance probabilities can be determined by simulating the distribution of a person-fit statistic T, for example l_z or l_z^*. This can be realized in at least two ways. One possibility is to determine the distribution of T by drawing a large number of θ-values from the standard normal distribution, each θ-value representing a person responding to the test. Then, item scores are simulated for each θ-value, according to the assumed IRT model and the CAT-procedure, and the value of T for each pattern is calculated; the T-values of these patterns constitute the simulated distribution based on the characteristics of the item bank. Another possibility is to simulate a distribution of T for each θ-value; that is, given θ, a large number of response patterns are simulated and for each pattern the T-value is calculated. So, now a distribution is simulated, based on the item bank and θ. Based on this distribution a critical value at level α can be determined.

The first method results in using one critical value for all simulees, whereas the second method will probably result in using different critical values at different θ-values. When the distribution of T is the same across all θ-levels, the first method will result in an appropriate simulated distribution.

Using the second method, the distribution of T can be simulated using a fixed sequence of items (test design) in which for each θ-value a large number of response vectors are simulated given the observed test design. Thus, each response vector consists of responses to the same items. However, an important aspect when simulating the distribution of a statistic is the stochastic process of item selection in a CAT (Glas, Meijer, & van Krimpen, 1997): in a CAT, the test design may be different for each simulee. To take this stochastic nature into account, the distribution can be simulated using a stochastic test design in which for each θ-value a large number of adaptive response patterns are simulated with, in principle, different test designs.

Let the vector $\mathbf{d}$ denote the test design, that is, a vector of the numbers of administered items in CAT, and $T(\mathbf{X})$ a statistic of the observed response vector $\mathbf{X} = (X_1, \cdots, X_k)$ of a test with k items. In CAT, item selection is based on responses to previous administered items which are dependent on the ability of the examinee. Therefore, $\mathbf{X}$ is conditional on $\mathbf{d}$ and θ,

and a function of $\mathbf{X}$, for example the statistic $T(\mathbf{X}) = T$, is also conditional on $\mathbf{d}$ and θ. Thus, the distribution of T, conditional on $\mathbf{d}$ and θ, is defined as

$$f(T|\,\mathbf{d},\theta)\,. \tag{10}$$

To obtain the unconditional distribution of T at a fixed θ-level, Equation 10 can be multiplied by the probability distribution of the design $\mathbf{d}$, which results in

$$f(T,\mathbf{d}\,|\theta) = f(\mathbf{d}|\,\theta)\,f(T|\,\mathbf{d},\theta)\,. \tag{11}$$

Comparing the values of the statistic across examinees with the same θ is difficult in a CAT environment, because in principle examinees respond to different tests. However, comparing significance probabilities of the observed value of the statistic across examinees is possible. For determining the significance probability, the distribution of a statistic, conditional on θ, can be simulated for a fixed test design (Equation 10) or a stochastic test design (Equation 11). In both approaches the distribution can be approximated by replicating the test n times.

Purpose of the Study

This study was designed to investigate (1) the distribution of l_z and l_z^* across different θ-levels and the influence of estimation errors of θ on the distribution of l_z and l_z^* for conventional testing (P&P) and in CAT (Study 1), (2) the influence of estimation errors of θ on simulating the distribution of l_0 and l_z for conventional tests and CAT, and the influence of the stochastic nature of the test design in CAT (Study 2), and (3) the detection rate of l_z and l_z^* for several types of aberrant response behavior in a conventional testing situation (Study 3). This study thus both extend the Nering (1997) and the Snijders (1998) study.

Study 1

In this study the distributions of l_z and l_z^* were investigated in a conventional and CAT situation. Nering (1997) examined the distribution of l_z in a CAT environment by first drawing $10,000$ θ-values from the standard normal distribution and then simulating adaptive response vectors for each θ-value. For each response vector, $\widehat{\theta}$ was used to determine the value of l_z. These $10,000$ values constituted the simulated distribution of l_z. In this study, the simulated distribution was determined by (1) drawing true θ from a standard normal distribution, or (2)

fixing true θ at different levels. In both (1) and (2), response vectors were generated and θ was estimated by $\widehat{\theta}$; $\widehat{\theta}$ was used to determine the value of l_z and l_z^*. Doing so, the critical values obtained by Nering can be compared with the critical values obtained when the distribution of l_z is simulated at a fixed θ-level. Finally, l_z was also calculated using true θ, that is $P(\theta)$ was used to determine the value of l_z. This enables us to investigate the influence of estimation errors in $\widehat{\theta}$ on the distribution of l_z. Note that l_z^* is only an appropriate standardization when $\hat{\theta}$ is used.

Method

P&P. Tests of 20, 50, and 80 items fitting the 2PLM were constructed, with $a_i \sim N(1; 0.2)$ and $b_i \sim U(-3; 3)$; each test was fixed for all simulees. For each test, ten datasets consisting of 10,000 response vectors were constructed. Nine datasets were constructed at nine different θ-levels: $\theta = -2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5$, and 2; one dataset was constructed where 10,000 θ's were drawn from a standard normal distribution.

First, for each response vector the values of l_z and l_z^* were calculated using $\widehat{\theta}$ and these 10,000 values of l_z and l_z^* were used to obtain the distribution of l_z and l_z^* for each dataset; θ was estimated using weighted maximum likelihood estimation (Warm, 1989); this estimator is less biased than the maximum likelihood estimator, and also exists for patterns with only 1-scores or only 0-scores. For all simulated distributions the critical values at level α were determined and compared with the critical values at level α of the standard normal distribution, where $\alpha = 0.01, 0.02, 0.03, 0.04$, and 0.05. Furthermore, the first four moments (mean, standard deviation, skewness, and kurtosis) of the simulated distribution of l_z and l_z^* were computed and compared with the moments of the standard normal distribution. Second, l_z was also calculated using true θ for each response vector to constitute the distribution of l_z without presence of estimation errors.

CAT. Ten datasets consisting of 10,000 adaptive response patterns were constructed. Nine datasets were constructed at nine different θ-levels: $\theta = -2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5$, and 2; one dataset was constructed where 10,000 θ's were drawn from a standard normal distribution.

A pool of 400 items fitting the 2PLM with $a_i \sim N(1; 0.2)$ and $b_i \sim U(-3; 3)$ was used. An adaptive response pattern was simulated as follows. First, the true θ of a simulee was drawn from a standard normal distribution or was set to a fixed θ-level, dependent on the dataset constructed. Then, the first item of the CAT selected was the item with maximum information given $\theta = 0$. For this item, $P(\theta)$, according to Equation 1 was determined. To simulate the

answer (1 or 0), a random number y from the uniform distribution on the interval $[0, 1]$ was drawn; when $y < P(\theta)$ the response to item i was set to 1 (correct response), 0 otherwise. The first four items of the CAT were selected with maximum information for $\theta = 0$, and based on the responses to these four items, $\widehat{\theta}$ was obtained. The next item selected was the item with maximum information given $\widehat{\theta}$. For this item, $P(\theta)$ was computed, a response was simulated, θ was estimated and another item was selected based on maximum information given $\widehat{\theta}$ at that stage. This procedure was repeated until the asymptotic standard error of $\widehat{\theta}$ was 0.25; this is an often used value, see for example DeAyala (1992) and Nering (1997). The asymptotic standard error of $\widehat{\theta}$ was determined by

$$SE\left(\widehat{\theta}\right) = \left[\sum_i a_i^2 P_i(\theta)\left(1 - P_i(\theta)\right)\right]^{-1/2}, \tag{12}$$

where the sum was across all administered items and $P_i(\theta)$ was defined by the 2PLM given in Equation 1; the standard error was estimated by substituting $\widehat{\theta}$ for θ.

For each response vector the values of l_z and l_z^* were calculated using $\widehat{\theta}$ and these 10,000 values of l_z and l_z^* were used to obtain the distribution of l_z and l_z^* for each dataset. Also, l_z was calculated using true θ for each response vector, to constitute the distribution of l_z without presence of estimation errors.

Results

Using $\widehat{\theta}$

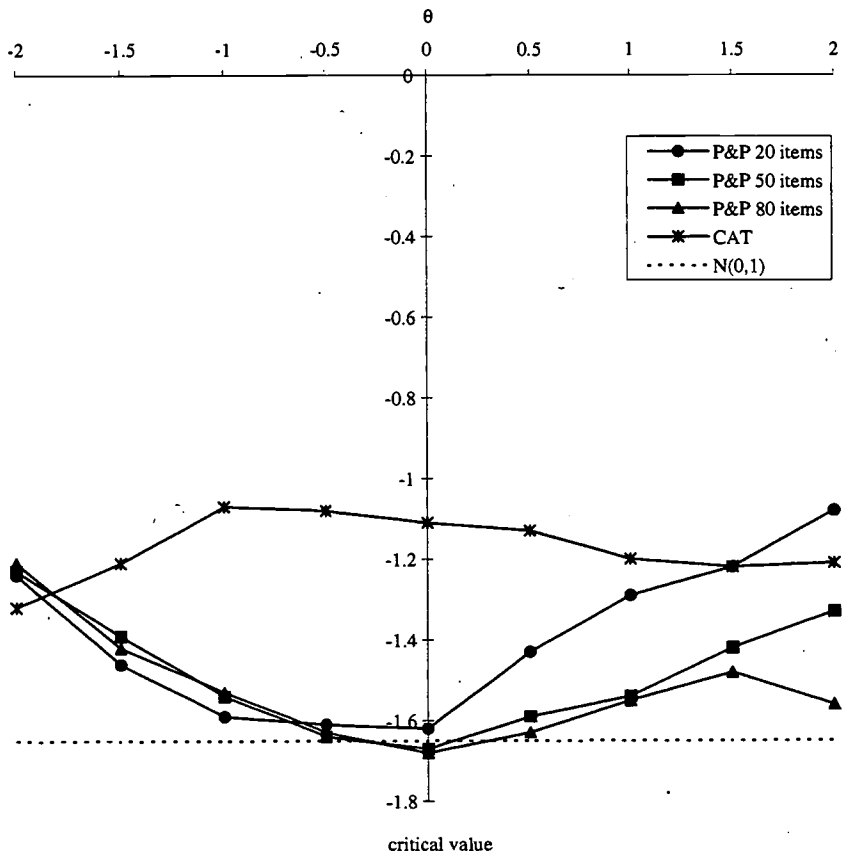
In Table 1 the first four moments and the critical values at level α of the simulated distribution of l_z, using $\widehat{\theta}$, are given for different θ-levels for a conventional test of 20, 50, and 80 items, and for a CAT. In Table 2 the first four moments and the critical values at level α of the simulated distributions of l_z^* are given for different θ-levels for a conventional test of 20, 50, and 80 items, and for a CAT.

P&P. Table 1 shows that, for the conventional test of 20 items, the mean and variance of the sampling distribution of l_z were different from 0 and 1 as expected under the standard normal distribution. For longer tests (50−80 items), the first two moments of the distribution of l_z were closer to the 0 and 1. However, the distribution tended to be negatively skewed, at all test lengths; for example, for the test with 80 items, the highest skewness observed was -0.36 for $\theta = -2.0$ and -1.5. The kurtosis tended to be slightly positive for all test lengths; for example for the test of length 50, the kurtosis varied between 0.10 and 0.45 for $\theta = -0.5$ and -2.0,

Table 1. Distributional characteristics of the simulated distribution of l_z, using $\hat{\theta}$.

						critical value				
		mean	variance	skewness	kurtosis	0.01	0.02	0.03	0.04	0.05
P&P 20 items										
$\theta \sim$	N(0,1)	0.16	0.77	-0.67	0.55	-2.25	-1.92	-1.73	-1.59	-1.46
$\theta =$	-2.0	0.16	0.56	-0.80	0.93	-1.98	-1.68	-1.47	-1.34	-1.24
	-1.5	0.12	0.75	-0.75	0.92	-2.34	-1.97	-1.74	-1.59	-1.46
	-1.0	0.12	0.88	-0.66	0.46	-2.55	-2.20	-1.94	-1.74	-1.59
	-0.5	0.12	0.94	-0.61	0.40	-2.53	-2.15	-1.94	-1.76	-1.61
	0.0	0.12	0.89	-0.69	0.58	-2.52	-2.16	-1.92	-1.77	-1.62
	0.5	0.17	0.78	-0.62	0.40	-2.27	-1.94	-1.72	-1.56	-1.43
	1.0	0.18	0.66	-0.68	0.52	-2.11	-1.77	-1.57	-1.41	-1.29
	1.5	0.20	0.58	-0.73	0.41	-1.90	-1.64	-1.46	-1.33	-1.22
	2.0	0.24	0.48	-0.85	0.75	-1.75	-1.45	-1.28	-1.16	-1.08
P&P 50 items										
$\theta \sim$	N(0,1)	0.09	0.86	-0.41	0.20	-2.32	-2.01	-1.80	-1.66	-1.54
$\theta =$	-2.0	0.09	0.54	-0.52	0.45	-1.86	-1.60	-1.43	-1.31	-1.23
	-1.5	0.09	0.69	-0.43	0.27	-2.07	-1.77	-1.61	-1.49	-1.39
	-1.0	0.08	0.87	-0.44	0.32	-2.43	-2.05	-1.82	-1.66	-1.54
	-0.5	0.09	0.97	-0.41	0.10	-2.51	-2.17	-1.92	-1.76	-1.64
	0.0	0.07	0.99	-0.42	0.22	-2.51	-2.20	-1.95	-1.80	-1.67
	0.5	0.09	0.94	-0.38	0.15	-2.42	-2.10	-1.87	-1.72	-1.59
	1.0	0.09	0.86	-0.37	0.13	-2.28	-2.01	-1.79	-1.65	-1.54
	1.5	0.10	0.76	-0.43	0.28	-2.21	-1.88	-1.68	-1.54	-1.42
	2.0	0.09	0.64	-0.46	0.33	-2.03	-1.74	-1.55	-1.43	-1.33
P&P 80 items										
$\theta \sim$	N(0,1)	0.06	0.89	-0.32	0.16	-2.37	-2.04	-1.84	-1.68	-1.58
$\theta =$	-2.0	0.08	0.55	-0.36	0.17	-1.87	-1.59	-1.43	-1.32	-1.21
	-1.5	0.07	0.71	-0.36	0.19	-2.10	-1.83	-1.67	-1.54	-1.42
	-1.0	0.07	0.85	-0.35	0.04	-2.28	-1.97	-1.78	-1.63	-1.53
	-0.5	0.06	0.97	-0.30	0.10	-2.45	-2.13	-1.91	-1.74	-1.63
	0.0	0.07	1.01	-0.34	0.16	-2.49	-2.16	-1.92	-1.79	-1.68
	0.5	0.08	0.96	-0.33	0.12	-2.47	-2.10	-1.90	-1.73	-1.63
	1.0	0.08	0.90	-0.31	-0.01	-2.28	-2.01	-1.82	-1.67	-1.55
	1.5	0.08	0.80	-0.33	0.14	-2.25	-1.95	-1.75	-1.58	-1.48
	2.0	0.07	0.91	-0.34	0.27	-2.45	-2.05	-1.83	-1.67	-1.56
CAT										
$\theta \sim$	N(0,1)	0.39	0.79	-0.20	0.03	-1.78	-1.51	-1.32	-1.22	-1.13
$\theta =$	-2.0	0.27	0.84	-0.36	0.06	-2.08	-1.77	-1.58	-1.44	-1.32
	-1.5	0.36	0.84	-0.25	0.07	-1.93	-1.65	-1.48	-1.32	-1.21
	-1.0	0.42	0.76	-0.17	-0.10	-1.70	-1.45	-1.27	-1.16	-1.07
	-0.5	0.40	0.75	-0.16	-0.19	-1.66	-1.44	-1.27	-1.17	-1.08
	0.0	0.40	0.77	-0.19	-0.11	-1.72	-1.48	-1.33	-1.21	-1.11
	0.5	0.41	0.81	-0.18	-0.05	-1.72	-1.50	-1.34	-1.23	-1.13
	1.0	0.38	0.83	-0.22	-0.10	-1.83	-1.60	-1.42	-1.29	-1.20
	1.5	0.34	0.82	-0.30	-0.01	-1.97	-1.64	-1.45	-1.33	-1.22
	2.0	0.32	0.78	-0.39	0.19	-1.99	-1.69	-1.48	-1.33	-1.21

Figure 1. Critical values, at $\alpha = 0.05$, of the simulated distribution of l_z, using $\hat{\theta}$.



respectively. A positive kurtosis indicates a leptokurtic distribution, that is, a distribution with heavier tails and a higher peak than the standard normal distribution. Table 1 also shows that for different θ-levels the observed critical values were different. For example, for a test of 20 items, the observed critical values at $\alpha = 0.01$ varied between -2.55 and -1.75 for $\theta = -1.0$ and 2.0, respectively, whereas the critical value at $\alpha = 0.01$ under the standard normal distribution is -2.33. In Figure 1 the critical values of the distribution of l_z, using $\widehat{\theta}$, at $\alpha = 0.05$, are plotted against θ for the conventional tests of 20, 50, and 80 items and a CAT. These critical values are compared with the critical value expected under the standard normal distribution, that is, -1.65. The distribution of l_z for a CAT will be discussed below. Figure 1 shows that for longer tests and for $-1 \leq \theta \leq 1$ the critical values observed in the simulated distribution were close to -1.65. So, especially for large positive and large negative θ-values the critical values in the simulated distribution were different from the expected critical value under the standard normal distribution.

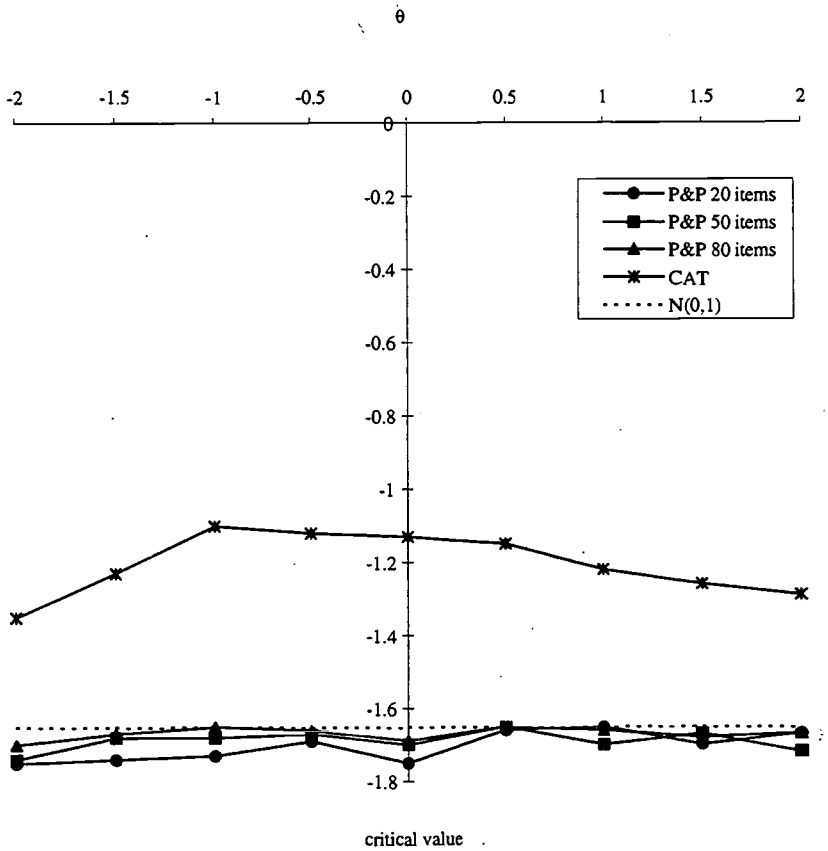
Table 2 shows that the mean and variance of the distribution of l_z^* were close to 0 and 1, respectively, for all θ-values and conventional tests of 20, 50 and 80 items. For example, for a test of 50 items the mean of l_z^* was 0.09 at $\theta = 0.5$ and $\theta = -0.5$, and 0.05 at $\theta = -2.0$, and the variance was for all θ-values approximately 1. However, for all tests and across all θ-values the distribution tended to be negatively skewed; for a test of 20 items, and $\theta = 0$ the skewness was -0.68. Also, the kurtosis tended to be positive; for example, a test of 80 items and $\theta = 0$ the kurtosis was 0.16. Table 2 also shows that the critical values in the simulated distribution were approximately equal across θ-levels. However, for smaller values of α the critical values in the simulated distribution differed from the critical values of the standard normal distribution. Moreover, for $\alpha \leq 0.04$ the use of l_z^* resulted in a slightly conservative classification of nonfitting response patterns; for example, for the standard normal distribution the critical value at $\alpha = 0.01$ is -2.33 and for the test of 50 items the critical value in the simulated distribution of l_z^* vary from -2.60 to -2.43 at $\theta = -1.0$, and $\theta = -1.5$, respectively. Figure 2 shows the critical values at $\alpha = 0.05$ of the simulated distribution of l_z^* across θ-levels for the three conventional tests and for a CAT. The results of the CAT will be discussed below. Table 2 and Figure 2 both show that the critical values at $\alpha = 0.05$ in the simulated distribution are close to -1.65 as expected for the standard normal distribution; for example, for the conventional test of 20 items the critical values at $\alpha = 0.05$ in the simulated distribution varied from -1.75 to -1.65 for $\theta = -2.0$ and $\theta = 1.0$, respectively.

CAT Environment. Table 1 shows that, for a CAT, the first two moments of the distribution of l_z are substantially different from 0 and 1 for all θ-levels; mean and variance

Table 2. Distributional characteristics of the simulated distribution of l_z^*.

						critical value				
		mean	variance	skewness	kurtosis	0.01	0.02	0.03	0.04	0.05
P&P 20 items										
$\theta \sim$	N(0,1)	0.13	0.97	-0.65	0.40	-2.53	-2.18	-1.97	-1.83	-1.67
$\theta =$	-2.0	0.09	0.98	-0.73	0.49	-2.68	-2.30	-2.01	-1.86	-1.75
	-1.5	0.10	1.01	-0.70	0.55	-2.73	-2.28	-2.05	-1.88	-1.74
	-1.0	0.11	1.00	-0.66	0.40	-2.72	-2.32	-2.07	-1.88	-1.73
	-0.5	0.11	1.01	-0.61	0.36	-2.64	-2.26	-2.02	-1.83	-1.69
	0.0	0.11	1.01	-0.68	0.51	-2.71	-2.30	-2.07	-1.88	-1.75
	0.5	0.15	0.98	-0.61	0.27	-2.54	-2.20	-2.00	-1.80	-1.66
	1.0	0.14	0.97	-0.67	0.37	-2.58	-2.25	-2.00	-1.82	-1.65
	1.5	0.14	0.97	-0.71	0.25	-2.60	-2.19	-2.00	-1.84	-1.70
	2.0	0.16	0.94	-0.77	0.39	-2.47	-2.16	-2.00	-1.82	-1.67
P&P 50 items										
$\theta \sim$	N(0,1)	0.08	0.99	-0.41	0.13	-2.50	-2.17	-1.94	-1.79	-1.67
$\theta =$	-2.0	0.05	1.00	-0.50	0.25	-2.60	-2.27	-2.04	-1.89	-1.74
	-1.5	0.07	0.98	-0.42	0.12	-2.43	-2.17	-1.96	-1.81	-1.68
	-1.0	0.07	1.01	-0.43	0.27	-2.60	-2.21	-2.00	-1.82	-1.68
	-0.5	0.09	1.01	-0.40	0.09	-2.59	-2.21	-1.97	-1.79	-1.67
	0.0	0.07	1.02	-0.42	0.21	-2.55	-2.23	-1.99	-1.82	-1.70
	0.5	0.09	1.01	-0.38	0.14	-2.52	-2.19	-1.95	-1.80	-1.65
	1.0	0.07	1.01	-0.38	0.12	-2.52	-2.18	-1.98	-1.82	-1.70
	1.5	0.08	1.01	-0.42	0.19	-2.56	-2.20	-1.97	-1.81	-1.67
	2.0	0.06	1.01	-0.46	0.23	-2.57	-2.22	-2.02	-1.86	-1.72
P&P 80 items										
$\theta \sim$	N(0,1)	0.05	1.00	-0.32	0.10	-2.52	-2.18	-1.97	-1.79	-1.68
$\theta =$	-2.0	0.05	0.99	-0.36	0.04	-2.54	-2.21	-2.00	-1.83	-1.70
	-1.5	0.05	0.99	-0.37	0.14	-2.55	-2.19	-1.98	-1.84	-1.69
	-1.0	0.07	0.98	-0.35	0.02	-2.43	-2.13	-1.93	-1.76	-1.65
	-0.5	0.06	1.01	-0.30	0.10	-2.49	-2.16	-1.94	-1.78	-1.66
	0.0	0.06	1.02	-0.34	0.16	-2.50	-2.17	-1.93	-1.80	-1.69
	0.5	0.07	1.00	-0.33	0.13	-2.52	-2.15	-1.93	-1.76	-1.65
	1.0	0.08	1.00	-0.31	-0.02	-2.42	-2.14	-1.94	-1.78	-1.66
	1.5	0.07	1.00	-0.33	0.10	-2.50	-2.18	-1.95	-1.79	-1.68
	2.0	0.06	1.01	-0.33	0.17	-2.57	-2.19	-1.95	-1.79	-1.67
CAT										
$\theta \sim$	N(0,1)	0.39	0.83	-0.23	0.10	-1.87	-1.56	-1.37	-1.27	-1.17
$\theta =$	-2.0	0.26	0.87	-0.39	0.14	-2.14	-1.84	-1.62	-1.48	-1.35
	-1.5	0.36	0.86	-0.28	0.14	-1.98	-1.68	-1.50	-1.35	-1.23
	-1.0	0.42	0.80	-0.19	-0.05	-1.77	-1.50	-1.32	-1.19	-1.10
	-0.5	0.41	0.80	-0.17	-0.17	-1.72	-1.49	-1.34	-1.21	-1.12
	0.0	0.40	0.79	-0.21	-0.06	-1.76	-1.54	-1.37	-1.23	-1.13
	0.5	0.41	0.82	-0.20	-0.00	-1.77	-1.53	-1.37	-1.25	-1.15
	1.0	0.38	0.85	-0.24	-0.04	-1.88	-1.64	-1.45	-1.32	-1.22
	1.5	0.34	0.86	-0.33	0.06	-2.05	-1.68	-1.50	-1.36	-1.26
	2.0	0.33	0.86	-0.43	0.29	-2.14	-1.82	-1.57	-1.42	-1.29

Figure 2. Critical values, at $\alpha = 0.05$, of the simulated distribution of l_z^*.



fluctuated around 0.40, and 0.80, respectively. Skewness and kurtosis were also different from 0. The distribution was found to be negatively skewed, and the highest skewness observed was -0.39 for $\theta = 2.0$. The highest kurtosis was found for $\theta = -0.5$ and 2.0 where the kurtosis was -0.19 and 0.19, respectively. Thus, the distribution of l_z using $\widehat{\theta}$ was quite different from the standard normal distribution. Table 1 and Figure 1 both show that the critical values in the sampling distribution tended to be closer to 0 than expected under the standard normal distribution for all θ and α. For example, the critical value at $\alpha = 0.05$ for $\theta = 0$, 5% of the simulees obtained a l_z-value below -1.11. Thus, using $l_z \leq -1.65$ will result in too few simulees being classified as aberrant; that is, the decision rule will result in a conservative classification of aberrant response behavior. Table 1 also shows that the critical values were different across θ-levels. For example, for $\theta = 0$ the critical value at $\alpha = 0.01$ was -1.72 whereas the critical value for $\theta = 2$ was -1.99. When θ was drawn from the standard normal distribution, the critical values were also closer to 0 than expected; for example, the critical value at $\alpha = 0.05$ was -1.13.

Table 2 shows that for a CAT the mean and variance of l_z^* were quite different from 0 and 1, respectively; for example, at $\theta = 0$ the mean and variance are 0.40 and 0.79, respectively. It also shows that the simulated distribution tended to be negatively skewed; the skewness varied from -0.17 to -0.43 at $\theta = -0.5$ and $\theta = -2.0$, respectively. The kurtosis was less systematically distributed; for $-1.0 \leq \theta \leq 1.0$ the kurtosis was slightly negative, for other θ-values positive kurtosis occurred. Figure 2 and Table 2 both show that the critical values in the simulated distribution of l_z^* were not in agreement with critical values of the standard normal distribution. For example, for $\alpha = 0.05$ the critical values in the simulated distribution varied from -1.10 to -1.35 at $\theta = -1.0$ and $\theta = -2.0$, respectively.

Using θ

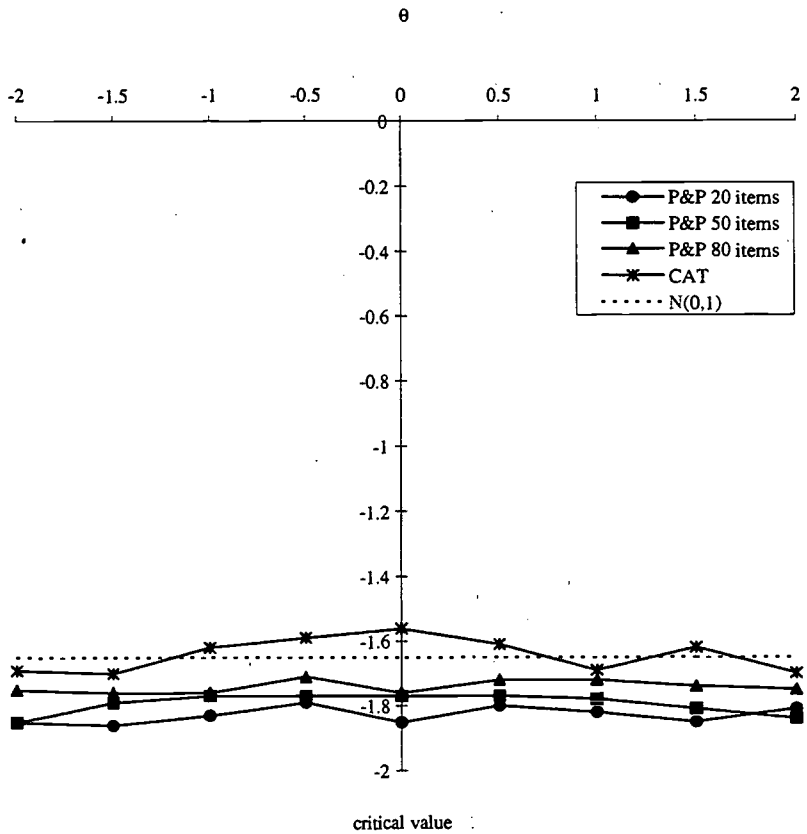
In Table 3 the first four moments and the critical values at level α of the simulated distributions of l_z, when true θ was used, are given.

P&P. Table 3 shows that, for all conventional tests, the first two moments of the distribution of l_z were close to 0 and 1, as expected under the standard normal distribution. However, the distributions are still negatively skewed and have positive kurtosis; for longer tests (50 – 80 items) the observed skewness fluctuated around 0.35 and the kurtosis around 0.11. Table 3 also shows that the critical values were about the same across θ-levels. However, the critical values tended to be slightly smaller than expected; for example the critical value at $\alpha = 0.03$ under the standard normal distribution is -1.88, and the values observed in the

Table 3. Distributional characteristics of the simulated distribution of l_z, using true θ.

						critical value				
		mean	variance	skewness	kurtosis	0.01	0.02	0.03	0.04	0.05
P&P 20 items										
$\theta \sim$	N(0,1)	0.02	0.98	-0.66	0.43	-2.73	-2.32	-2.09	-1.93	-1.78
$\theta =$	-2.0	-0.01	0.98	-0.77	0.74	-2.78	-2.39	-2.16	-2.00	-1.85
	-1.5	-0.00	1.03	-0.77	0.75	-2.87	-2.49	-2.12	-2.01	-1.86
	-1.0	0.01	1.01	-0.67	0.41	-2.80	-2.42	-2.22	-1.99	-1.83
	-0.5	0.01	1.01	-0.56	0.32	-2.75	-2.34	-2.11	-1.92	-1.79
	0.0	-0.01	1.00	-0.63	0.50	-2.76	-2.41	-2.18	-1.99	-1.85
	0.5	0.02	0.98	-0.63	0.36	-2.66	-2.34	-2.08	-1.92	-1.80
	1.0	0.01	0.99	-0.70	0.38	-2.75	-2.36	-2.14	-1.95	-1.82
	1.5	-0.00	0.99	-0.75	0.42	-2.73	-2.42	-2.17	-1.99	-1.85
	2.0	0.01	0.97	-0.96	1.20	-2.90	-2.47	-2.20	-1.96	-1.81
P&P 50 items										
$\theta \sim$	N(0,1)	0.00	0.99	-0.42	0.10	-2.54	-2.25	-2.05	-1.89	-1.76
$\theta =$	-2.0	-0.02	1.02	-0.58	0.38	-2.76	-2.36	-2.13	-1.97	-1.85
	-1.5	0.00	1.01	-0.51	0.24	-2.67	-2.32	-2.11	-1.92	-1.79
	-1.0	-0.00	1.01	-0.47	0.30	-2.71	-2.34	-2.09	-1.90	-1.77
	-0.5	0.01	1.02	-0.42	0.12	-2.63	-2.29	-2.10	-1.90	-1.77
	0.0	-0.01	1.02	-0.40	0.19	-2.64	-2.32	-2.06	-1.90	-1.77
	0.5	-0.00	1.02	-0.38	0.12	-2.63	-2.27	-2.03	-1.90	-1.77
	1.0	-0.01	0.99	-0.39	0.07	-2.53	-2.17	-2.01	-1.90	-1.78
	1.5	-0.01	1.02	-0.53	0.34	-2.73	-2.35	-2.13	-1.94	-1.81
	2.0	-0.02	1.00	-0.61	0.42	-2.82	-2.39	-2.20	-2.01	-1.84
P&P 80 items										
$\theta \sim$	N(0,1)	-0.01	1.01	-0.34	0.12	-2.60	-2.24	-2.04	-1.88	-1.75
$\theta =$	-2.0	-0.00	0.97	-0.48	0.30	-2.64	-2.27	-2.06	-1.90	-1.75
	-1.5	-0.00	0.98	-0.43	0.11	-2.66	-2.27	-2.03	-1.87	-1.76
	-1.0	-0.00	0.99	-0.39	0.13	-2.58	-2.23	-2.02	-1.86	-1.76
	-0.5	-0.00	1.01	-0.32	0.11	-2.57	-2.23	-2.03	-1.87	-1.71
	0.0	-0.00	1.02	-0.33	0.16	-2.58	-2.23	-2.00	-1.88	-1.76
	0.5	0.00	1.00	-0.33	0.13	-2.58	-2.22	-2.00	-1.85	-1.72
	1.0	0.01	1.00	-0.32	0.00	-2.51	-2.24	-2.00	-1.85	-1.72
	1.5	-0.00	1.00	-0.41	0.21	-2.62	-2.29	-2.05	-1.89	-1.74
	2.0	-0.01	1.01	-0.36	0.23	-2.68	-2.29	-2.04	-1.88	-1.75
CAT										
$\theta \sim$	N(0,1)	0.04	0.95	-0.23	0.06	-2.38	-2.10	-1.87	-1.72	-1.62
$\theta =$	-2.0	0.01	0.95	-0.36	0.05	-2.51	-2.16	-2.00	-1.83	-1.69
	-1.5	0.01	0.99	-0.24	0.01	-2.47	-2.16	-1.95	-1.81	-1.70
	-1.0	0.05	0.95	-0.24	0.05	-2.39	-2.09	-1.86	-1.74	-1.62
	-0.5	0.05	0.94	-0.22	-0.03	-2.36	-2.06	-1.87	-1.71	-1.59
	0.0	0.06	0.92	-0.28	0.10	-2.35	-2.03	-1.83	-1.67	-1.56
	0.5	0.05	0.95	-0.20	0.01	-2.30	-2.03	-1.84	-1.72	-1.61
	1.0	0.02	0.97	-0.28	0.04	-2.43	-2.13	-1.93	-1.81	-1.69
	1.5	0.04	0.93	-0.32	0.02	-2.36	-2.10	-1.88	-1.73	-1.62
	2.0	0.02	0.94	-0.40	0.16	-2.53	-2.21	-2.00	-1.84	-1.70

Figure 3. Critical values, at $\alpha = 0.05$, of the simulated distribution of l_z, using θ.



simulated distributions for $k = 80$ are close to -2.00. In Figure 3 the critical values at $\alpha = 0.05$, of the conventional tests and a CAT, are plotted against θ and compared with the critical value expected under the standard normal distribution, that is -1.65. Figure 3 shows that the critical values are approximately the same across θ-levels and that the critical values using simulated data have larger negative values than expected under the standard normal distribution.

CAT Environment. Using θ to determine the distribution of l_z resulted in a mean and variance close to 0 and 1, as expected under the standard normal distribution. However, the distribution tended to be negatively skewed, with the largest value of -0.40 for $\theta = 2.0$. For $\theta = 2.0$ the highest kurtosis of 0.16 was obtained. It can be concluded that the distribution of l_z using true θ was more in agreement with the standard normal distribution than when $\widehat{\theta}$ was used. Similar conclusions pertain for the critical values. Table 3 and Figure 3 both show that the critical values were close to -1.65 as expected under the standard normal distribution.

Study 2

In Study 1 it was shown that the critical values of the distribution of l_z^* were close to the critical values of the standard normal distribution for conventional tests. It was also shown that for long conventional tests ($50 - 80$ items) and $-1 \leq \theta \leq 1$ the critical values of l_z were reasonably in agreement with the standard normal distribution. However, for extreme positive and negative θ-levels, the critical values found in the simulated distribution were quite different than expected under the standard normal distribution. An alternative to using critical values from the theoretical distribution is to simulate a distribution for a person-fit statistic for each simulee. In this second study, the distributions of l_0 and l_z are simulated for every simulee, and the influence of estimation errors of θ on the distributions of l_0 and l_z were investigated in a conventional testing and CAT environment.

With respect to CAT, it was shown in Study 1 that (1) the distributions of l_z and l_z^* did not follow a standard normal distribution and (2) the distributions of l_z and l_z^* differed across θ-levels when $\widehat{\theta}$ was used; as a result, it is advisable to simulate the distribution conditional on θ or $\widehat{\theta}$. Another aspect of this second study was to investigate the influence of the stochastic nature of the test design in CAT.

Method

P&P. Eight datasets of 400 model fitting response vectors fitting the 2PLM were constructed with $a_i \sim N(1, 0.2)$ and $b_i \sim U(-3, 3)$; each dataset contained a test of different

test length, and each test was fixed for all simulees. Test length was $k = 10, 20, 30, 40, 50, 60, 70,$ and, 80 items. True θ was drawn from the standard normal distribution, where each θ-value represented a simulee responding to a test; θ was estimated by $\widehat{\theta}$ using, weighted maximum likelihood estimation (Warm, 1989). For each simulee, the distributions of l_0 and l_z were simulated in two different ways, both using parametric bootstrap techniques (Efron, 1982). First, for each simulee it was assumed that θ equalled $\widehat{\theta}$. For example, suppose a simulee with true parameter value $\theta = 1.5$ responded to a test and $\widehat{\theta} = 1.2$; then, for each simulee, $1,000$ replications were generated with $\widehat{\theta} = 1.2$. For each replicated response pattern the values of l_0 and l_z were determined to obtain the simulated distribution; $\widehat{\theta}$ was used to compute the value of l_0 and l_z. Then, the values of l_0 and l_z of the original response patterns, also computed using $\widehat{\theta}$, were compared with the simulated distribution by determining the significance probability under the sampled distribution. Second, it was assumed that the true parameter value was θ; for example, for a simulee with true parameter $\theta = 1.5$ and $\widehat{\theta} = 1.2$, it was assumed that the true parameter value was known and was 1.5. For each simulee, $1,000$ replications were generated where the known true parameter value was set to θ. Doing this, estimation errors in θ are excluded from approximating the distribution of l_0 and l_z. For each replicated response pattern the values of l_0 and l_z were determined using true θ to obtain the simulated distribution. Then, the values of l_0 and l_z of the original response patterns, also computed using true θ, were compared with the simulated distribution by determining the significance probability under the sampling distribution. Also, for each dataset the mean absolute bias was determined as $MAB = \frac{1}{n}\sum_1^n \left|\widehat{\theta} - \theta\right|$, where the sum is across all simulees.

Note, that for conventional tests the distribution of l_0 and l_z are equivalent; the distribution is simulated conditional on θ or $\hat{\theta}$, all items are the same, and therefore, for every replication $E(l_0)$ and var (l_0) are the same.

CAT. A dataset of 400 model fitting adaptive response patterns was constructed using the item pool and procedure described in Study 1, where true θ was drawn from the standard normal distribution. The distributions of l_0, l_z, and l_z^* were simulated in two different ways, both using parametric bootstrap methods. First, for each adaptive response vector these distributions were simulated using a fixed test design (cf. Equation 10). For each simulee, 500 response patterns were replicated where the test design was set to the observed test design. Thus, for each simulee, the administered test was viewed as a *conventional test*, and this conventional test was replicated 500 times, conditional on the value of θ or $\hat{\theta}$; the values of l_0 and l_z were computed for each replicated response pattern to obtain the distribution given θ or $\widehat{\theta}$ and d, whereas the values of l_z^* were only determined using $\hat{\theta}$. Then, the significance probability was

determined by comparing the values of l_0, l_z, and l_z^* of the original response pattern with the simulated distribution.

Second, the distributions of l_0, l_z, and l_z^* were simulated using the stochastic test design (cf. Equation 11). For each simulee, an *adaptive test* was replicated 500 times, where true θ or $\widehat{\theta}$ was used; the values of l_0 and l_z were computed using θ or $\widehat{\theta}$, whereas the values of l_z^* were only computed using $\hat{\theta}$. For each simulee, 500 adaptive response patterns given $\widehat{\theta}$ or θ were replicated according to the CAT-procedure described in Study 1; that is, $P(\theta)$ or $P\left(\widehat{\theta}\right)$ was used to generate responses to items. This procedure was repeated until $SE\left(\widehat{\theta}\right) \leq 0.25$. Thus, for each simulee 500 adaptive response patterns were simulated conditional on the value of θ or $\widehat{\theta}$. For each replicated response pattern the values of l_0, l_z, and l_z^* were determined to obtain the simulated distribution. Then, the values of l_0, l_z, and l_z^* of the original response patterns were compared with the simulated distribution by determining the significance probabilities. When the replications were generated using $\widehat{\theta}$ all l_0, l_z, and l_z^* values were determined by using $\widehat{\theta}$ in Equations 2 and 3. When the replications were replicated using θ, all l_0 and l_z values were computed using θ in order to determine a distribution without the presence of estimation errors. Although in practice θ is unknown, determining the distribution based on θ allow us to investigate the influence of $\widehat{\theta}$ on the distribution of l_0 and l_z.

Note, that for the fixed design the distribution of l_0, l_z, and l_z^* are equivalent; the distribution is simulated conditional on θ or $\hat{\theta}$, all items are the same due to the fixed test design, and therefore, for every replication $E(l_0)$ and var (l_0) are the same.

Results

P&P. In Table 4 the distribution of the significance probabilities of l_z are given, using θ or $\widehat{\theta}$ to determine the values of l_z. To illustrate the distribution of the significance probabilities, ten intervals are considered, each of length 0.10. The expected proportion of simulees with a significance probability in a particular interval was 0.10; for example, it was expected that 10% of the simulees have l_z-values with $0.4 < p* \leq 0.5$. To test whether the distribution of significance probabilities approached the uniform distribution Pearson's chi-squared tests, X^2, can be calculated, with $E(X^2) = 9$. Table 4 shows that, for all test lengths, the distribution of significance probabilities are uniformly distributed. However, in practice θ is unknown and as an alternative $\hat{\theta}$ is used. Table 4 shows that, when $\hat{\theta}$ was used, for very short tests, containing only 10 items, the significance probabilities are not uniformly distributed ($X^2 = 24.4$, p-value= 0.04). For tests containing 20 items or more, the distribution of l_z, using $\hat{\theta}$, is more in agreement with the uniform distribution. Table 4 also shows that, in general, for small values

Table 4. Distribution of significance probabilities for a P&P test of k items, using and $\hat{\theta}$ to simulate the null distribution of l_0 or l_z.

		Significance probabilities in interval										
k	MAB	[0,0.1]	(0.1,0.2]	(0.2,0.3]	(0.3,0.4]	(0.4,0.5]	(0.5,0.6]	(0.6,0.7]	(0.7,0.8]	(0.8,0.9]	(0.9,1.0]	X^2
θ												
10	0.64	.095	.125	.103	.083	.090	.085	.110	.095	.103	.113	6.3
20	0.46	.103	.113	.098	.085	.118	.100	.085	.095	.088	.118	5.7
30	0.39	.115	.120	.105	.113	.090	.103	.073	.103	.095	.085	7.7
40	0.32	.110	.095	.090	.083	.105	.110	.108	.100	.098	.103	2.9
50	0.29	.145	.098	.088	.105	.108	.093	.108	.080	.078	.100	13.0
60	0.27	.105	.098	.078	.088	.113	.115	.070	.103	.100	.133	12.2
70	0.26	.125	.068	.153	.085	.103	.108	.098	.098	.085	.080	21.5
80	0.26	.103	.105	.100	.085	.080	.118	.088	.103	.105	.115	5.5
$\hat{\theta}$												
10	0.64	.058	.080	.075	.118	.093	.090	.125	.128	.138	.098	24.4
20	0.46	.095	.083	.083	.093	.095	.090	.130	.108	.100	.125	9.6
30	0.39	.055	.115	.103	.128	.098	.095	.125	.103	.098	.083	16.0
40	0.32	.078	.085	.085	.103	.085	.125	.108	.120	.120	.093	10.9
50	0.29	.115	.075	.088	.123	.105	.080	.110	.118	.078	.110	11.8
60	0.27	.080	.085	.083	.110	.120	.095	.083	.100	.125	.120	11.2
70	0.26	.108	.080	.130	.090	.108	.100	.113	.100	.075	.098	9.2
80	0.26	.078	.083	.113	.093	.088	.120	.098	.113	.108	.110	7.2

Note: 400 simulees and 1000 replications per simulee

of the MAB, the distribution can be approximated using simulation methods based on $\widehat{\theta}$. For example, for the short test of 10 items and using $\hat{\theta}$, $MAB = 0.64$ and $X^2 = 24.4$ whereas for the test of 80 items $MAB = 0.26$ and $X^2 = 7.2$.

CAT. In Table 5 the distribution of the significance probabilities of l_0, l_z, and l_z^* using a fixed and a stochastic design are given using θ and $\widehat{\theta}$. Table 5 shows that, using θ and using a fixed or stochastic test design resulted in an approximately uniform distribution of the significance probabilities for both l_0 and l_z. However, conditioning on $\widehat{\theta}$ and using a fixed or stochastic test design resulted in an inappropriate approximation of the distribution of l_0, l_z and l_z^*. For example, using $\widehat{\theta}$ and a stochastic test design with l_z as fit index resulted in $X^2 = 65.7$, which is highly significant. Especially the probabilities in the left tail were much too small. Using a stochastic design, only 4.3% of the simulees attained a l_z-value with $0 \leq p* \leq 0.1$. In practice, using $\widehat{\theta}$ and a stochastic test design to simulate the distribution performs better than a using fixed test design. That is, the values of X^2 are lower when the stochastic test design was used compared with using a fixed design; for l_0 the values of X^2 for a fixed and stochastic test design were 81.7 and 34.8, respectively, for l_z 81.7 and 65.7, and for l_z^* 81.7 and 65.7, respectively.

Study 3

In Study 1 it was shown that the critical value at $\alpha = 0.05$ of the distribution of l_z^* was close to -1.65. This third study was designed to compare the detection rate of l_z^* with l_z for several types of aberrant response behavior, when it is assumed that the theoretical distribution is standard normal.

Method

Several datasets containing 200 nonfitting response patterns were constructed, with three types of aberrant response behavior and three different conventional tests; tests containing 10, 20 and 50 items. The first type was guessing on all the items in a test. This guessing model mimics the type of answering behavior studied empirically by Van den Brink (1977). He described examinees who took a multiple choice exam without preparation, and the only purpose of taking the exam was to become familiar with the type of questions that would be asked. Because returning an almost completely blank answering sheet may focus attention on an examinee's ignorance, the examinee would randomly guess the correct answer on almost all items in the test. "Guessing" simulees were simulated by randomly guessing the correct answer

Table 5. Distribution of significance probabilities for a CAT, using a fixed or stochastic design, using θ and $\hat{\theta}$,to simulate the null distribution of l_0, l_z, or l_z^*.

		Significance probabilities in interval										
	test design	[0,0.1]	(0.1,0.2]	(0.2,0.3]	(0.3,0.4]	(0.4,0.5]	(0.5,0.6]	(0.6,0.7]	(0.7,0.8]	(0.8,0.9]	(0.9,1.0]	X^2
θ												
l_0	fixed	0.093	0.088	0.098	0.088	0.098	0.100	0.110	0.110	0.100	0.118	3.6
	stochastic	0.095	0.093	0.095	0.090	0.105	0.105	0.115	0.080	0.115	0.108	4.7
l_z	fixed	0.093	0.088	0.098	0.088	0.098	0.100	0.110	0.110	0.100	0.118	3.6
	stochastic	0.098	0.093	0.105	0.098	0.088	0.098	0.103	0.100	0.120	0.100	2.7
$\hat{\theta}$												
l_0	fixed	0.030	0.065	0.063	0.048	0.088	0.140	0.133	0.123	0.140	0.173	81.7
	stochastic	0.058	0.075	0.055	0.080	0.110	0.120	0.128	0.108	0.123	0.145	34.8
l_z	fixed	0.030	0.065	0.063	0.048	0.088	0.140	0.133	0.123	0.140	0.173	81.7
	stochastic	0.043	0.060	0.063	0.058	0.095	0.138	0.120	0.123	0.133	0.170	65.7
l_z^*	fixed	0.030	0.065	0.063	0.048	0.088	0.140	0.133	0.123	0.140	0.173	81.7
	stochastic	0.033	0.065	0.063	0.058	0.103	0.135	0.125	0.128	0.130	0.163	65.7

Note: 400 simulees and 500 replications per simulee

on each item with a probability of 0.2 (assuming a test with five alternatives per item).

Second, response vectors with a two-dimensional θ parameter were simulated: a simulee had during the first half of the test another ability value than during the second half to respond to the items. Carelessness, fumbling or memorization of some items can be the cause of non-invariant abilities. Two datasets containing response vectors with a two-dimensional ability parameter were simulated by drawing two ability values, θ_1 and θ_2, from a bivariate standard normal distribution; the correlation between the two values was modeled by the parameter ρ. Thus, during the first half of the test $P(\theta_1)$ was used and during the second half $P(\theta_2)$ was used to simulate the responses to the items. The values $\rho = 0.8$ and $\rho = 0.6$ were used here to simulate the response patterns.

The third type of aberrant response vectors simulated were vectors with violations against local stochastic independence between the items of the test. When previous items provide new insights that are useful for answering the next item, or when the process of answering the items is exhausting, the assumption of local independence between the items may be violated. Four datasets were constructed with violations of the local independence assumption. These response vectors were simulated according to a model proposed by Jannarone (1986). Appendix B describes the model in detail. Using this model, the probability of correctly answering an item is now determined by the item parameters a and b, the person parameter θ and the association parameter δ. When $\delta = 0$ the model equals the 2PLM. Compared to the 2PLM, positive values of δ result in a higher probability of a correct response, and negative values of δ result in a lower probability of correctly answering an item. The values $\delta = -2, -1, 1$, and 2 were used to simulate these nonfitting response patterns.

The detection rate of a statistic is defined here as the proportion of detected nonfitting response patterns. A response vector was classified as nonfitting the 2PLM when the observed value of l_z or l_z^* was below the critical value at level $\alpha = 0.05$ of the standard normal distribution, that is -1.65.

For every dataset, the mean absolute bias, MAB, was determined and the mean bias was calculated as $MB = \frac{1}{n}\sum_1^n \left(\widehat{\theta} - \theta\right)$. These variables were determined to investigate the trade off between bias and detection rate.

Results

In Table 6 the detection rates of l_z and l_z^* are given for three types of aberrant response behavior, for conventional tests of 10, 20, and 50 items. Table 6 shows that for all types of aberrance and all tests the detection rate of l_z^* was slightly higher than the detection rate of

Table 6. Detection rates for several types of aberrant response behavior, for P&P-tests of length 10, 20 and 50, using $l_z \leq -1.65$ and $l_z^* \leq -1.65$.

		k=10			k=20			k=50		
		MAB	l_z	l_z^*	MAB	l_z	l_z^*	MAB	l_z	l_z^*
guessing		2.27	0.12	0.27	2.25	0.45	0.65	2.04	0.85	0.96
$\rho =$	0.6	0.74	0.05	0.06	0.60	0.04	0.07	0.42	0.06	0.07
	0.8	0.83	0.03	0.07	0.53	0.07	0.08	0.38	0.04	0.06
$\delta =$	-2.0	0.91	0.05	0.07	0.83	0.07	0.09	0.71	0.18	0.20
	-1.0	0.69	0.03	0.05	0.55	0.06	0.09	0.46	0.04	0.06
	1.0	1.09	0.01	0.04	0.79	0.03	0.05	0.58	0.04	0.06
	2.0	1.53	0.02	0.03	1.41	0.05	0.09	1.23	0.06	0.07

l_z. For example, for simulees guessing on all items of a conventional test of length 20, the detection rates for l_z^* and l_z were 0.65 and 0.45, respectively. Table 6 also shows that for guessing the detection rates were reasonably high, whereas for violations of local independence and unidimensionality of θ, the detection rates for both l_z and l_z^* were low for all test lengths. For example, for 50 items and guessing the detection rates were 0.85 and 0.96 for l_z and l_z^*, respectively, whereas for violation of unidimensionality and $\rho = 0.6$ the detection rates were 0.06 and 0.07 for l_z and l_z^*, respectively. Table 6 also shows that the relation between MAB and detection rate was unclear. For example, for guessing on all items on the 50 items test, the MAB was high (2.04) and the detection rates were also high (0.85 and 0.96 for l_z and l_z^*, respectively). However, for violation of local independence and $\delta = 2.0$ and 10 items, the MAB was rather high (1.53) but the detection rates were low (0.02 and 0.03 for l_z and l_z^*, respectively).

Discussion

To detect examinees with inappropriate test scores, the use of person-fit statistics was investigated in this study. In particular the distribution using theoretical and simulated distributions, and using θ and $\widehat{\theta}$ in a conventional and CAT environment were explored. In Study 1 the empirical distributions of l_z and l_z^* were compared with the theoretical distribution (i.e., standard normal) for conventional and adaptive tests. Results showed that, for conventional tests, the distribution of l_z differed across θ-levels. However, for θ-values between -1 and 1 and long tests ($50 - 80$ items), the critical values at $\alpha = 0.05$ of the simulated distribution (using $\hat{\theta}$ to determine the values of l_z) were close to the expected -1.65 (see Reise, 1995, for similar findings in the context of personality assessment). The critical values at $\alpha = 0.05$ of the empirical distribution of l_z^* for conventional tests and for all θ-values were found to be close to -1.65, as expected under the standard normal distribution. With respect to CAT, results showed that the distribution of both l_z and l_z^* differed across θ-levels and that the critical values of the theoretical distribution differed substantial from the critical values of the empirical distributions using $\widehat{\theta}$.

In Study 2, simulating the distributions of l_0, l_z, and l_z^* to create an approximation of the empirical distribution for conventional and adaptive tests was investigated. In a conventional testing context, especially for large positive and large negative θ-values, simulating a sampling distribution of l_z for every examinee based on $\hat{\theta}$ resulted in an appropriate approximation of the distribution. With respect to CAT, simulating the distributions of l_0, l_z, and l_z^* was problematic. For all three statistics, the left tails of the simulated distribution were inaccurate; for example,

using a stochastic test design to simulate the distribution of l_0 for every simulee resulted in only 5.8% of the simulees attaining a value of l_0 in the left 10% area of the distribution.

In Study 3 the detection rate of l_z and l_z^* to detect nonfitting response patterns was investigated in a conventional testing context. Results showed that l_z^* performed slightly better than l_z for short tests. For long tests the differences in detection rates between l_z and l_z^* were smaller because for long tests the critical values of the empirical distribution of l_z were reasonably in agreement with the critical values of the standard normal distribution (see also Study 1).

A possible solution for the problems in simulating the sampling distribution may be to use Bayesian methods to 'lift up' the tails of the distribution. Other alternatives that may be considered in future research are using less biased estimators of θ, or using statistics that are less dependent on $\hat{\theta}$.

Appendix A. Derivation of l_z^*

Snijders (1998) derived the asymptotic distribution of statistics which are linear in the item responses, and in which θ was replaced by an estimate. Statistics are linear in the item response when the statistic can be written as

$$\sum_{i=1}^{n} X_i w_i(\theta) - w_0(\theta), \tag{13}$$

where $w_i(\theta)$ are suitable functions. Snijders used in his paper the centered form

$$W_n(\theta) = \sum_{i=1}^{n} (X_i - P_i(\theta)) w_i(\theta). \tag{14}$$

For example, $W_n = l_0 - E(l_0)$ and $w_i = \ln \frac{P_i(\theta)}{1-P_i(\theta)}$ results in the centered version of l_0. The only restriction on the estimator $\hat{\theta}$ was, that $\hat{\theta}$ satisfied an equation of the form

$$r_0\left(\hat{\theta}\right) + \sum_{i=1}^{n} \left(X_i - P_i\left(\hat{\theta}\right)\right) r_i\left(\hat{\theta}\right) = 0. \tag{15}$$

For example, for the maximum likelihood estimator and the 2PLM, $r_0\left(\hat{\theta}\right) = 0$ and $r_i\left(\hat{\theta}\right) = a_i$ for $i = 1, \cdots, n$.

The estimate satisfying

$$\frac{J\left(\hat{\theta}\right)}{2I\left(\hat{\theta}\right)} + \sum_{i=1}^{n} \left[X_i - P_i\left(\hat{\theta}\right)\right] \frac{P_i'\left(\hat{\theta}\right)}{P_i\left(\hat{\theta}\right)\left[1 - P_i\left(\hat{\theta}\right)\right]} = 0, \tag{16}$$

where

$$I\left(\hat{\theta}\right) = \sum_{i=1}^{n} \frac{P_i'^2\left(\hat{\theta}\right)}{P_i\left(\hat{\theta}\right)\left[1 - P_i\left(\hat{\theta}\right)\right]} \tag{17}$$

$$J\left(\hat{\theta}\right) = \sum_{i=1}^{n} \frac{P_i'\left(\hat{\theta}\right) P_i''\left(\hat{\theta}\right)}{P_i\left(\hat{\theta}\right)\left[1 - P_i\left(\hat{\theta}\right)\right]}, \text{ and} \tag{18}$$

$$P_i'(\theta) = \partial P_i(\theta)/\partial\theta,\ P_i''(\theta) = \partial^2 P_i(\theta)/\partial\theta^2. \tag{19}$$

is the Warm estimator. Therefore, after some algebra, for the 2PLM, $r_i = a_i$ and

$$r_0\left(\hat{\theta}\right) = \frac{\sum_{i=1}^{n} a_i^3 P_i\left(\hat{\theta}\right)\left[1 - P_i\left(\hat{\theta}\right)\right]\left[1 - 2P_i\left(\hat{\theta}\right)\right]}{2\sum_{i=1}^{n} a_i^2 P_i\left(\hat{\theta}\right)\left[1 - P_i\left(\hat{\theta}\right)\right]}. \tag{20}$$

Snijders showed that the expected value of $W_n\left(\hat{\theta}\right)$ can be approximated by

$$E\left(W_n\left(\hat{\theta}\right)\right) \approx -c_n\left(\hat{\theta}\right) r_0\left(\hat{\theta}\right), \tag{21}$$

and the variance by

$$var\left(W_n\left(\hat{\theta}\right)\right) \approx n\tau_n^2\left(\hat{\theta}\right), \tag{22}$$

where

$$\tau_n^2\left(\hat{\theta}\right) = \frac{1}{n}\sum_{i=1}^{n} \tilde{w}_i^2\left(\hat{\theta}\right) P_i\left(\hat{\theta}\right)\left[1 - P_i\left(\hat{\theta}\right)\right], \tag{23}$$

$$\tilde{w}_i\left(\hat{\theta}\right) = w_i\left(\hat{\theta}\right) - c_n\left(\hat{\theta}\right) r_i\left(\hat{\theta}\right), \text{ and} \tag{24}$$

$$c_n\left(\hat{\theta}\right) = \frac{\sum_{i=1}^{n} P_i'\left(\hat{\theta}\right) w_i\left(\hat{\theta}\right)}{\sum_{i=1}^{n} P_i'\left(\hat{\theta}\right) r_i\left(\hat{\theta}\right)}. \tag{25}$$

He also showed that the asymptotic distribution of

$$\frac{W_n\left(\hat{\theta}\right) + c_n\left(\hat{\theta}\right) r_0\left(\hat{\theta}\right)}{\sqrt{n}\tau_n\left(\hat{\theta}\right)} \tag{26}$$

is standard normal. Note that the value of $c_n\left(\hat{\theta}\right) r_0\left(\hat{\theta}\right)$ does not depend directly on the patterns of item responses, but only on $\hat{\theta}$.

Appendix B. Modeling Local Dependence

Let x_i, a realization of X_i, be the response to item i. One of the models Jannarone (1986) presented was a conjunctive Rasch-model; local dependence between two subsequent items can be modeled by

$$P(X_i = x_i, X_{i+1} = x_{i+1} \mid \theta) = \frac{\exp\left[\sum_{j=i}^{i+1} x_j(\theta - b_j) + x_i x_{i+1}\left(\theta - \xi_{i,i+1}\right)\right]}{1+\exp[\theta - b_i]+\exp[\theta - b_{i+1}]+\exp\left[\sum_{j=i}^{i+1}(\theta - b_j)+\left(\theta - \xi_{i,i+1}\right)\right]}, \quad (27)$$

where $\xi_{i,i+1}$ is a parameter modeling association between items i and $i+1$. This model can be generalized to a conjunctive 2PLM, which can be written as

$$P(X_i = x_i, X_{i+1} = x_{i+1} \mid \theta) \propto \exp\left[\sum_{j=i}^{i+1} x_j a_j (\theta - b_j) + x_i x_{i+1}\gamma_{i,i+1}\left(\theta - \xi_{i,i+1}\right)\right], \quad (28)$$

where γ and ξ are parameters modeling association between items.

In this study, the following model was used to simulated response vectors with local independence between all subsequent items was

$$P(X_i = x_i, X_{i+1} = x_{i+1} \mid \theta) \propto \exp\left[\sum_{j=i}^{i+1} x_j a_j (\theta - b_j) + x_i x_{i+1}\delta_{i,i+1}\right], \quad (29)$$

where $\delta_{i,i+1}$ is a parameter modeling association between items. The four possible realizations of (X_i, X_{i+1}) have the following probabilities

$$\begin{aligned}
P(X_i = 0, X_{i+1} = 0) &\propto 1, \\
P(X_i = 1, X_{i+1} = 0) &\propto \exp\left[a_i(\theta - b_i)\right], \\
P(X_i = 0, X_{i+1} = 1) &\propto \exp\left[a_{i+1}(\theta - b_{i+1})\right], \text{ and} \\
P(X_i = 1, X_{i+1} = 1) &\propto \exp\left[a_i(\theta - b_i) + a_{i+1}(\theta - b_{i+1}) + \delta_{i,i+1}\right].
\end{aligned}$$

The conditional probability of a correct response to item $i+1$ given a correct response to item i can be written as

$$\begin{aligned}
P(X_{i+1} = 1 \mid X_i = 1, \theta) &= \frac{P(X_i=1, X_{i+1}=1|\theta)}{P(X_i=1, X_{i+1}=0|\theta)+P(X_i=1, X_{i+1}=1|\theta)} \\
&= \frac{\exp[a_i(\theta - b_i)+a_{i+1}(\theta - b_{i+1})+\delta_{i,i+1}]}{\exp[a_i(\theta - b_i)]+\exp[a_i(\theta - b_i)+a_{i+1}(\theta - b_{i+1})+\delta_{i,i+1}]}, \quad (30)
\end{aligned}$$

and the probability of a correct response to item $i+1$ given an incorrect response to the previous item can be written as

$$P\left(X_{i+1}=1\,|X_i=0,\theta\right)=\frac{\exp[a_{i+1}(\theta-b_{i+1})]}{1+\exp[a_{i+1}(\theta-b_{i+1})]}, \tag{31}$$

which is the 2PLM. The conditional probabilities in Equation 30 and the 2PLM were used to simulate the responses to the items when the items are local stochastic dependent.

Author Note

This study received funding from the Law School Admission Council (LSAC). The opinions and conclusions contained in this report are those of the authors and not necessarily reflect the position or policy of LSAC.

The names of the authors are alphabetical; they are equal responsible for the contents of this paper.

References

Baker, F. B. (1992). *Item response theory: Parameter estimation techniques.* New York: Dekker.

DeAyala, R. J. (1992). The nominal response model in computerized adaptive testing. *Applied Psychological Measurement, 16*, 327-343.

Drasgow, F. & Levine, M. V. (1986). Optimal detection of certain forms of inappropriate test scores. *Applied Psychological Measurement, 10*, 59-67.

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

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

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

Drasgow, F., Levine, M. V., & Zickar, M. J. (1996). Optimal identification of mismeasured individuals. *Applied Measurement in Education, 9(1)*, 47-64.

Glas, C. A. W., Meijer R. R. & van Krimpen, E. M. L. A. (1997). *Statistical tests for person misfit in computerized adaptive testing.* Research Report 97-08, University of Twente, Enschede.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans.* Philadelphia, Pa: Society for Industrial and Applied Mathematics (SIAM).

Hambleton, R. K., & Swaminatan, H. (1985). *Item response theory: Principles and applications* (2nd ed.). Boston: Kluwer-Nijhoff Publishing.

Jannarone, R. J. (1986). Conjunctive item response theory kernels. *Psychometrika, 51*, 357-373.

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

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

Meijer, R. R. & Nering, M. L. (1997). Trait level estimation for nonfitting response vectors. *Applied Psychological Measurement, 21*, (321-336).

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

Psychometrika, 55, 75-106.

Nering, M. L. (1996). The effects of person misfit in computersized adaptive testing (Doctoral dissertation, University of Minnesota). *Dissertation Abstracts International, 57*, 04B.

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

Reise, S. P. (1995). Scoring method and the detection of person misfit in a personality assessment context. *Applied Psychological Measurement, 19*, 213-229.

Reise, S. P. & Waller (1990). Fitting the two-parameter model to personality data. *Applied Psychological Measurement, 14*, 45-58.

Snijders, T. (1998). *Asymptotic Distribution of Person-Fit Statistics with Estimated Person Parameter.* Unpublished report, University of Groningen, The Netherlands.

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

Van den Brink (1977). Het verken-effect [The scouting effect]. *Tijdschrift voor Onderwijsresearch, 2*, 253-261.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427-450.

Zickar, M. J., & Drasgow, F. (1996). Detecting faking on a personality instrument using appropriateness measurement. *Applied Psychological Measurement, 20*, 71-87.

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

RR-98-02	R.R. Meijer & E.M.L.A. van Krimpen-Stoop, *Simulating the Null Distribution of Person-Fit Statistics for Conventional and Adaptive Tests*
RR-98-01	C.A.W. Glas, R.R. Meijer, E.M.L.A. van Krimpen-Stoop, *Statistical Tests for Person Misfit in Computerized Adaptive Testing*
RR-97-07	H.J. Vos, *A Minimax Sequential Procedure in the Context of Computerized Adaptive Mastery Testing*
RR-97-06	H.J. Vos, *Applications of Bayesian Decision Theory to Sequential Mastery Testing*
RR-97-05	W.J. van der Linden & Richard M. Luecht, *Observed-Score Equating as a Test Assembly Problem*
RR-97-04	W.J. van der Linden & J.J. Adema, *Simultaneous Assembly of Multiple Test Forms*
RR-97-03	W.J. van der Linden, *Multidimensional Adaptive Yesting with a Minimum Error-Variance Criterion*
RR-97-02	W.J. van der Linden, *A Procedure for Empirical Initialization of Adaptive Testing Algorithms*
RR-97-01	W.J. van der Linden & Lynda M. Reese, *A Model for Optimal Constrained Adaptive Testing*
RR-96-04	C.A.W. Glas & A.A. Béguin, *Appropriateness of IRT Observed Score Equating*
RR-96-03	C.A.W. Glas, *Testing the Generalized Partial Credit Model*
RR-96-02	C.A.W. Glas, *Detection of Differential Item Functioning using Lagrange Multiplier Tests*
RR-96-01	W.J. van der Linden, *Bayesian Item Selection Criteria for Adaptive Testing*
RR-95-03	W.J. van der Linden, *Assembling Tests for the Measurement of Multiple Abilities*
RR-95-02	W.J. van der Linden, *Stochastic Order in Dichotomous Item Response Models for Fixed Tests, Adaptive Tests, or Multiple Abilities*
RR-95-01	W.J. van der Linden, *Some decision theory for course placement*
RR-94-17	H.J. Vos, *A compensatory model for simultaneously setting cutting scores for selection-placement-mastery decisions*
RR-94-16	H.J. Vos, *Applications of Bayesian decision theory to intelligent tutoring systems*
RR-94-15	H.J. Vos, *An intelligent tutoring system for classifying students into Instructional treatments with mastery scores*
RR-94-13	W.J.J. Veerkamp & M.P.F. Berger, *A simple and fast item selection procedure for adaptive testing*
RR-94-12	R.R. Meijer, *Nonparametric and group-based person-fit statistics: A validity study and an empirical example*

RR-94-10 W.J. van der Linden & M.A. Zwarts, *Robustness of judgments in evaluation research*

RR-94-9 L.M.W. Akkermans, *Monte Carlo estimation of the conditional Rasch model*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Research Reports can be obtained at costs, Faculty of Educational Science and Technology, University of Twente, Mr. J.M.J. Nelissen, P.O. Box 217, 7500 AE Enschede, The Netherlands.

faculty of

EDUCATIONAL SCIENCE AND TECHNOLOGY

A publication by
The Faculty of Educational Science and Technology of the University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

NOTICE

REPRODUCTION BASIS

This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").