Measurement and Research Department Reports 2002-1

Testing the unidimensionality assumption of the Rasch model

Norman Verhelst

Cito Arnhem

Measurement and Research Department Reports 2002-1

Testing the unidimensionality assumption of the Rasch model

Norman Verhelst

Citogroep
Arnhem, Januari 2001

Abstract

Statistical tests especially designed to test the unidimensionality axiom of the Rasch model are scarce. For two of them, the Martin-Löf test (ML-test) and the splitter-item-technique, an extensive power analysis has been carried out, showing clearly the superiority of the ML-test. The disadvantage of the ML-test, however, is that its null distribution deviates strongly from the asymptotic chi-square distribution unless one has huge samples. A new test with one degree of freedom is proposed. Its power is superior to that of the ML-test, and its null distribution converges rapidly to the chi-square.

Introduction

The assumption of unidimensionality is at the heart of the Rasch model and of many other IRT-models as well. Nonetheless, relatively few attention has been paid to the statistical testing of this assumption Tests especially sensitive to violation of the unidimensionality axiom are scarce. For parametric statistical tests of the Rasch model, aimed at detection of multidimensionality, the list is very short. All there is can be related to two important contributions, developed from quite different viewpoints.

The first and oldest contribution is a test developed by Martin-Löf (1973, see also Gustafsson, 1980, Verhelst, 1993, and Glas and Verhelst, 1995). To apply the test, it is assumed that the dimensional composition is known, the items fall into two subsets, such that all items in the same subset represent the same dimension.

The other contribution is Van den Wollenberg's Q_2-test (1979, 1982), where the alternative hypothesis is quite vague. The author has shown that the test has power against violation of the unidimensionality assumption. He conjectured that the asymptotic null distribution of the Q_2-test statistic is the chi square distribution, but a proof was never given. Thorough theoretical work by Glas (1989) showed that the conjecture of Van den Wollenberg is probably not correct. Fortunately, Glas was able to derive a modification of the Q_2-test statistic (called the R_{2c}-statistic), which is asymptotically chi-square distributed. The computation of this test statistic, however, is utterly complicated, and in practice only feasible if the number of items in the test is small.

A third approach, also initiated by van den Wollenberg (o.c.), and elaborated by Molenaar (1983) is the so-called splitter-item-technique. In this approach one item (the splitter item) is used as a criterion test. The sample is split according

to the score on this splitter item, and the parameters of the remaining items are estimated (using conditional maximum likelihood, CML) in each of the two samples. The product of the two conditional likelihoods is then compared with the likelihood after estimation of the parameters in the total sample by means of a likelihood ratio test. Van den Wollenberg (1979) has shown that this test has power against violation of the unidimensionality assumption.

This ends the list of tests especially constructed to detecting multidimensionality. For all tests listed, the asymptotic distribution has been shown to be chi-square (except for the Q_2-test), and for all, some evidence of their power has been demonstrated. A systematic study of the power of these tests, and a comparison of their power curves, however, has never been undertaken.

The primary purpose of the present article is to explore the power of these tests and to present a comparison of their power curves as a function of some measure of deviation of unidimensionality. It was readily realized, however, that the power of tests as the ones mentioned here depends on many factors, and that consequently, any attempt to quantify the impact of many of these factors might result in a disordered and probably chaotic collection of tables, from which it may be very hard to extract recommendations for practical applications. Therefore, a number of restrictions has been imposed throughout. Here is a list:

1. All violations to the unidimensionality assumption have the same structure: the latent variable is a bivariate normally distributed variable (θ_1, θ_2) with zero means and unit variances in both dimensions. The severity of the violation is expressed by the correlation ρ between the two variates. The test consists of k items, which can be partitioned into two sets of k_1 and k_2 items respectively. For both subsets the Rasch model holds with θ_1 and θ_2, respectively, as latent variable. The Rasch

model then holds for the set of k items if and only if $\rho = 1$. All power curves will be reported as a function of the correlation.

2. In all studies reported, all item parameters are equal to zero. Although the power of all statistical tests considered depends on the distribution of the item parameters, there seems to be no reason to expect that the comparison of the power curves of the tests will be influenced in an essential way by altering the distribution of the difficulty parameters.
3. Although sample size is the most important tool to manipulate the power of a statistical test, we did not vary the sample size in the power studies to be reported, but kept it constant at a value of 1000. We have no reason to assume that the differences between the power curves of the tests we considered, will change in important respects for other sample sizes, except maybe for small values, where the approximation of the null distribution by the asymptotic distribution becomes problematic anyway, but this aspect of the tests is not investigated thoroughly, although it will be discussed at two places.
4. The significance level of all tests is set at 5%.

The set up of a power study like the present one is very simple in principle: for a given test and a given value of ρ, thousand artificial response patterns are generated, the test statistic is computed and the binary result (significant or not) is recorded. This procedure is repeated a number of times, and the proportion of significant results is an estimate of the power at the used value of ρ. In all tables to be presented the number of replications is 4000, except in the following case. The power is always estimated for a series of values of ρ, starting at 1 and decreasing in steps of 0.05 until $\rho = 0.5$. If for a certain value of ρ the estimated power exceeded 99%, the estimate for all subsequent values was set at 100% without further computations.

While developing the specialized software to carry out the analyses, it was decided not to study the Q_2- or R_{2c}-tests. The Q_2 was excluded because its asymptotic null distribution is not known, and the theoretically better founded R_{2c} is available. But this test is hardly applicable in tests having more than 15 items, so that the programming effort will almost never be put at use in practice. While studying the theoretical similarities and differences between the Martin-Löf test and the R_{2c}-test, however, we discovered that it was fairly easy to construct a class of statistical tests with high power against violation of unidimensionality and having only one degree of freedom. This class of tests will be discussed at some length in the sequel, and their power will be compared with that of the Martin-Löf test and the splitter-item-technique.

The Martin-Löf test

Theoretical considerations

The Martin-Löf test for multidimensionality (ML-test) can be conceived of as a likelihood ratio test. By hypothesis, the item set under consideration is split into two or more subsets and the likelihood is maximized under the general hypothesis that the Rasch model (RM) is valid for each subset. The null hypothesis or restricted model states that the RM is valid for all items jointly.

To investigate the power of the ML-test, we will stick to the case of two subsets, having k_1 and k_2 items respectively. The subsets themselves will be denoted by their index sets I_1 and I_2. Response patterns will be denoted as $\mathbf{x}_1$ and $\mathbf{x}_2$, and test scores as s_1 and s_2. The corresponding symbols without subscript will denote the corresponding quantities for all items jointly: $k = k_1 + k_2$, $s = s_1 + s_2$ and $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$. The symbol π will denote the probability function.

Martin-Löf considers the marginal likelihood of a score pattern under the general model ($\pi(\mathbf{x}_1, \mathbf{x}_2)$), which can be written as

$$\begin{aligned}\pi(\mathbf{x}_1, \mathbf{x}_2) &= \pi(\mathbf{x}_1, \mathbf{x}_2|s_1, s_2) \times \pi(s_1, s_2) \\ &= \pi(\mathbf{x}_1|s_1)\ \pi(\mathbf{x}_2|s_2) \times \pi(s_1, s_2), \end{aligned} \tag{1}$$

where the '$\times$'-sign indicates a factorization of the likelihood in a conditional part and a marginal part. For the marginal part ML uses a saturated multinomial model. The item parameter estimates are completely determined by the conditional part.

Now if we take the logarithm of (1), and sum it across all observations, we will get a sum due to the conditional part and a sum due to the marginal part. Using the tilde to denote the maximum under the general model, we find

$$\ln \widetilde{L}_g = \ln \widetilde{L}_{gc} + \ln \widetilde{L}_{gm}, \tag{2}$$

where

$$\ln \widetilde{L}_{gc} = \sum_{i \in I_1} t_i \ln(\widetilde{\varepsilon}_i) + \sum_{i \in I_2} t_i \ln(\widetilde{\varepsilon}_i) - \sum_{s_1} n_{s_1} \ln(\gamma_{s_1}(\widetilde{\boldsymbol{\varepsilon}}_1)) - \sum_{s_2} n_{s_1} \ln(\gamma_{s_2}(\widetilde{\boldsymbol{\varepsilon}}_2)), \tag{3}$$

and

$$\ln \widetilde{L}_{gm} = \sum_{s_1} \sum_{s_2} n_{s_1 s_2} \ln(n_{s_1 s_2}/n), \tag{4}$$

where ε_i is the easiness parameter of item i, expressed on an exponential scale, t_i is the number of correct responses to item i, n_{s_ℓ} is the observed number of response patterns $\mathbf{x}_\ell$ having a score equal to s_ℓ ($\ell = 1,2$), $n_{s_1 s_2}$ is the observed number of response patterns $(\mathbf{x}_1, \mathbf{x}_2)$ with score (s_1, s_2), and $\gamma_s(\boldsymbol{\varepsilon})$ denotes the basic symmetric

function of order s and with a vector $\boldsymbol{\varepsilon}$ as argument:

$$\gamma_s(\boldsymbol{\varepsilon}) = \sum_{\sum x_i = s} \prod_i \varepsilon_i^{x_i}.$$

Under the restricted model, the marginal likelihood of a response pattern can be written as

$$\pi(\mathbf{x}_1, \mathbf{x}_2) = \pi(\mathbf{x}|s) \times \pi(s), \tag{5}$$

and the restricted log-likelihood at its maximum can be written as

$$\ln \widehat{L}_r = \sum_i t_i \ln(\widehat{\varepsilon}_i) - \sum_s n_s \ln(\gamma_s(\widehat{\boldsymbol{\varepsilon}})) + \sum_s n_s \ln(n_s/n). \tag{6}$$

Notice that in (3) and in (6) the sufficient statistics for the item parameters, t_i, are identical.

In the Rasch model the conditional probability of obtaining the scores s_1 and s_2 on two exclusive subtests, given the total score $s = s_1 + s_2$ is given by

$$\pi(s_1, s_2|s) = \frac{\gamma_{s_1}(\boldsymbol{\varepsilon}_1)\gamma_{s_2}(\boldsymbol{\varepsilon}_2)}{\gamma_s(\boldsymbol{\varepsilon})}, \tag{7}$$

such that the maximum likelihood estimator of the cell proportion in the bivariate score distribution under the restricted model is given by

$$\frac{n_s}{n} \frac{\gamma_{s_1}(\widehat{\boldsymbol{\varepsilon}}_1)\gamma_{s_2}(\widehat{\boldsymbol{\varepsilon}}_2)}{\gamma_s(\widehat{\boldsymbol{\varepsilon}})}. \tag{8}$$

Now, consider the following sum of two terms:

$$A = \sum_{s_1} n_{s_1} \ln(\gamma_{s_1}(\widehat{\boldsymbol{\varepsilon}}_1)) + \sum_{s_2} n_{s_2} \ln(\gamma_{s_2}(\widehat{\boldsymbol{\varepsilon}}_2)), \tag{9}$$

then we find by adding A to (6) and subtracting again that

$$\ln \widehat{L}_r = \ln \widehat{L}_{gc} + \ln \widehat{L}_{gm}, \tag{10}$$

meaning that we can write the restricted log-likelihood in the same formal way as the unrestricted one. The only thing that we have to do is to use in the unrestricted case the estimates $\widetilde{\varepsilon}$, and in the restricted case the estimates $\widehat{\varepsilon}$ (and using (8) to evaluate $\ln \widehat{L}_{gm}$).

The Martin-Löf test statistic can then be written as

$$ML = 2\left[\left(\ln \widetilde{L}_{gc} - \ln \widehat{L}_{gc}\right) + \left(\ln \widetilde{L}_{gm} - \ln \widehat{L}_{gm}\right)\right]. \tag{11}$$

As one can see, the ML-statistic consists of two terms, both of which are positive. The first term compares the conditional log-likelihood of the general model, evaluated at the maximum under the general model and the restricted model respectively. The second term reflects the comparison between the observed bivariate frequencies $n_{s_1 s_2}$ ($\ln \widetilde{L}_{gm}$) and the predicted bivariate frequencies under the restricted model ($\ln \widehat{L}_{gm}$).

An interesting observation is the following. In all the power studies done with the ML-statistic, the relative contribution of each of the two terms in (11) has been recorded. The contribution of the first term was never larger than 2.5%. Contributions larger than 0.05% were only observed for small k (< 16); for longer tests the contribution of the first term was negligible for all values of ρ considered. It follows that for practical purposes one might safely use an approximation to the ML-statistic:

$$ML \approx ML^* = 2\left(\ln \widetilde{L}_{gm} - \ln \widehat{L}_{gm}\right), \tag{12}$$

and since the computation of $\ln \widetilde{L}_{gm}$ is trivially simple (see (4)), this approximation

means a considerable reduction in the work to be done for the computations: the item parameters have to be estimated only under the restricted model, which is the RM for all items jointly. Since the ignored term is positive, use of the approximation will yield a slightly conservative test. The computational formula for the approximation is

$$ML^* = 2\sum_{s_1,s_2} n_{s_1,s_2} \ln\left[\frac{n_{s_1,s_2}}{n_s}\frac{\gamma_s(\widehat{\varepsilon})}{\gamma_{s_1}(\widehat{\varepsilon}_1)\gamma_{s_2}(\widehat{\varepsilon}_2)}\right]. \tag{13}$$

The degrees of freedom for the ML-test is

$$df(ML) = k_1k_2 - 1. \tag{14}$$

In the sequel, all power studies on the ML-test are based on (11), and not on its approximation (13).

The power of the ML-test

To get an impression of the power of the ML-test, three series of analyses were carried out, following the general setup described in the Introduction. In each series the two dimensions are represented by an equal number of items ($k_1 = k_2$), taking the values 3, 5 and 8 respectively. The results of the study are displayed in Table 1. The graphs of the power curves are given in Figure 1

Three remarks will be made with respect to these results. First, there is a very marked effect of the test length on the power of the test. Second, the rejection rate for a test with 16 items when the null hypothesis is true ($\rho = 1$) is suspiciously low. In the next subsection detailed attention to this phenomenon will be given. The third aspect of the results is the disappointingly low power, even with 16 items, for correlations which are very common in cognitive testing. In the PISA project

Table 1. Power of the ML-test (in %)

ρ	$k_1 = k_2 = 3$	$k_1 = k_2 = 5$	$k_1 = k_2 = 8$
1.0	5.33	5.78	2.55
.95	6.50	8.78	6.98
.90	11.1	18.8	26.1
.85	21.0	39.4	62.8
.80	33.7	64.9	90.6
.75	53.4	86.9	98.9
.70	69.8	96.3	100.
.65	83.8	99.3	100.
.60	92.1	100.	100.
.55	97.1	100.	100.
.50	99.1	100.	100.

(OECD, 2001) a correlation in the order of magnitude of .85 was found between dimensions as distinct as reading and mathematics. For more similar dimensions (like subscales of mathematics) the correlations are usually as high as 0.9 (frequently found in the Dutch National Assessment Program). It should be remembered that these correlations are not attenuated by unreliability: they are correlations between latent variables, not between observed scores.

The null distribution in the ML-test

As can be seen from Table 2, the rejection rate in case the null hypothesis is true ($\rho = 1$) is far too low for the case $k_1 = k_2 = 8$. Verguts and De Boeck (2000) offer as a possible explanation that the expected frequencies in the bivariate frequency table are too small for too many cells. As an example, the expected frequencies under the null hypothesis have been computed for a sample of $10,000$ observations with $k_1 = 8$ and $k_2 = 11$ items respectively in the two test halves. The results are displayed in Table 2. As an extreme low score in one test halve and an extreme high score in the

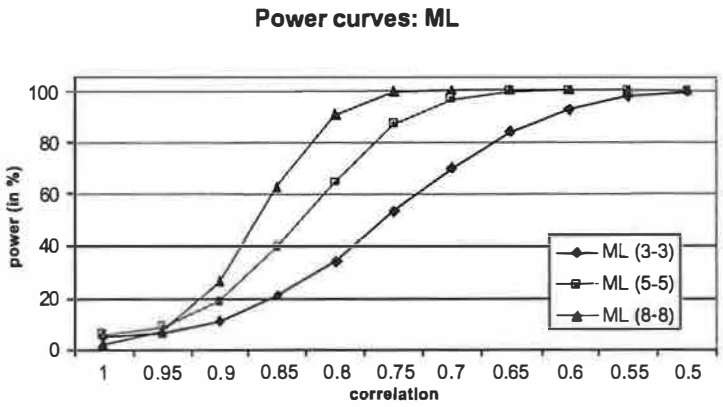

Figure 1

other test halve is highly unlikely, the corresponding expected frequencies are very low, and this might make the null distribution of the ML-test statistic systematically deviant from the chi-square distribution. Of the $9 \times 12 = 108$ cells in the table, 22 have an expected frequency smaller than 5.

To investigate the form of the true null distribution, the ML-test statistic was computed (under the null hypothesis) on $10,000$ samples of $1,000$ and $10,000$ samples of $10,000$ observations , and the percentiles $(1, \ldots, 99)$ were determined in both empirical distributions and compared to the percentiles in the chi-square distribution with 87 degrees of freedom. The results are displayed as Q-Q-plots in Figure 2, from which it is very clear that the null distribution in the ML-test deviates systematically from the chi-square distribution in the sense that the empirical percentiles are systematically smaller than the chi-square percentiles. The deviation is more pronounced for $n = 1000$ than for $n = 10,000$. At the nominal rejection rate of 5%, the null hypothesis is rejected only in 1.5% of the cases for $n = 1000$. To get a rejection rate of 5%, one should not use the critical value under the chi-square distribution (which is 109.773), but the critical value under the exact null distrib-

Table 2. Expected frequencies under the null hypothesis ($n = 10,000$)

$s_2 \backslash s_1$	0	1	2	3	4	5	6	7	8
0	73.7	78.3	51.2	24.4	9.5	2.9	0.7	0.1	0.0
1	107.7	161.0	134.4	83.6	39.8	15.1	4.4	0.8	0.1
2	100.7	192.0	209.0	159.3	94.4	43.9	14.8	3.6	0.5
3	72.0	179.1	239.0	226.7	164.7	88.9	38.3	10.9	1.6
4	44.8	136.6	226.7	263.6	222.3	153.1	76.5	25.4	4.7
5	23.9	90.7	184.5	249.0	267.9	214.3	124.5	52.7	11.3
6	11.3	52.7	124.5	214.3	267.9	249.0	184.5	90.7	23.9
7	4.7	25.4	76.5	153.1	222.3	263.6	226.7	136.6	44.8
8	1.6	10.9	38.3	88.9	164.7	226.7	239.0	179.1	72.0
9	0.5	3.6	14.8	43.9	94.4	159.3	209.0	192.0	100.7
10	0.1	0.8	4.4	15.1	39.8	83.6	134.4	161.0	107.7
11	0.0	0.1	0.7	2.9	9.5	24.4	51.2	78.3	73.7

ution, which is in this example is estimated at 100.852 This means that, at least in this case, the ML-test is conservative, and as a consequence that the power is underestimated. Our finding is in line with the findings of Verguts and De Boeck, and can probably be generalized to all instances of the ML-test.

The splitter-item-technique

Rationale and examples

The term splitter-item-technique was introduced by Molenaar (1983), in an elaboration of an approach introduced by Van den Wollenberg (1979). We will paraphrase here the rationale as proposed by Van den Wollenberg (1979, pp. 108-112).

Suppose a test measures two dimensions, and we have an external criterion which measures one of these dimensions. We split the sample of testees according to this criterion in a high scoring group and a low scoring group. The items loading on the criterion related dimension will be relatively difficult for the low group and

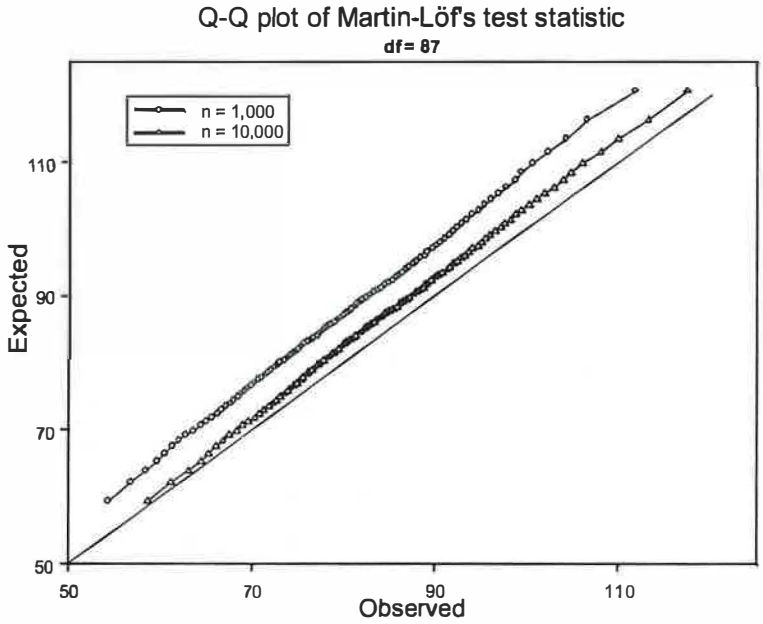

Figure 2.

relatively easy for the high group, while the relative difficulty of the other items will be less affected by the criterion variable: not if the dimensions are independent, and more so as the correlation between the two dimensions grows. Building a likelihood ratio test using this splitting of the sample will have power against the violation of unidimensionality, because the rank order of the difficulties will tend to differ in both subsamples.

Using such an approach to compare with the ML-test, of course, is not fair, because in the ML-test we do not have an external criterion. To make a fair comparison we should only rely on the test data. We can do this if we use some of the items as a criterion measure, and build the test on the other items. If only one item is used, the technique is called the splitter-item-technique (Molenaar, 1983), but we could use more items. Using more items will generally lead to a more valid splitting

of the sample, but at the same time will lessen the number of items to be used in the test, and thus to decreasing power. So there is a trade-off between the validity of the splitting and the test length for the remaining items on which the test is built.

To refer to the whole class of tests based on this rationale we will use the general term 'likelihood ratio' test (LR) and reserve the term splitter-item-technique to the particular case where one item is used as a criterion variable.

The power of such a test will, however, depend on more features than just the number of items used as splitting criterion. We name a few of them: (i) Molenaar found that using the splitter-item-technique where the splitting is done on a very easy or very difficult item does not work well. Probably because the two subsamples are of very unequal size. (ii) The splitting can be done in two groups, but more groups can be used of course, allowing for more opportunities to show differences in difficulty, but at the same time paying in increasing the number of degrees of freedom. It is our guess that two groups (preferably of equal size) will maximize power. (iii) Not only the number of remaining test items will count, but also their balance. We expect that a more balanced test will show more power than an unbalanced test.

In the general setup of this report, we do not intend to investigate the effect of the difficulty of the criterion item(s) on the power, but the effect of the other factors may be explored by using a quite simple design. The number of items used as a criterion will be denoted k_c, and power is investigated for the six cases displayed in Table 3.

To allow for a comparison with the ML-test, the power was estimated for $8+11 = 19$ items and the rejection rate was computed at the nominal 5% level ($df = 87$, critical chi-square value equals 109.773), as well as at the real 5% level (critical value: 100.852; see the preceding subsection). The results are displayed in Table 4.

Table 3. Design of power study

case	k_1	k_2	k_c	#groups	df
1	8	10	1	2	17
2	7	11	1	2	17
3	8	8	3	2	15
4	8	8	3	4	45
5	5	11	3	2	15
6	5	11	3	4	45

The numbers 1 to 6 in the top row refer to the six cases described in Table 3.

It is seen from Table 4 that the ML-test outperforms all six versions of the LR-test, except for a correlation of 0.95 if the chi square approximation is used. The differences between the power curves for the six cases will be discussed in some more detail in the next subsection.

Table 4. Power Analysis Results

ρ	ML-real	ML-nom.	1	2	3	4	5	6
1	5.00	1.55	5.10	4.90	5.32	5.50	4.92	5.25
.95	13.5	5.02	6.12	6.60	6.72	6.87	6.65	6.57
.90	46.6	25.8	9.50	8.92	13.2	10.6	11.5	10.6
.85	83.9	68.2	17.5	16.0	27.5	21.8	24.4	19.3
.80	98.7	95.2	28.4	28.0	47.1	39.6	41.4	32.0
.75	99.9	99.7	43.2	43.2	69.4	60.7	62.7	53.5
.70	100.	100.	61.5	58.3	85.3	80.2	79.7	72.6
.65	100.	100.	75.3	73.9	94.2	92.5	91.1	87.7
.60	100.	100.	85.6	84.6	98.4	97.8	96.4	95.7
.55	100.	100.	92.4	92.4	99.5	99.5	99.0	98.9
.50	100.	100.	96.6	96.3	100.	100.	100.	100.

Summary of the power study

From Table 4, it is clear that the ML-test is more powerful than the LR-tests. In Figure 3 power curves are displayed for four cases: the ML-test with rejection at the nominal level and at the real level (diamonds with plain lines and dashed lines respectively), and the two cases using the splitter-item-technique (squares representing a more balanced test (case 1) than the triangles (case 2)). The more balanced case (having 8 and 10 items respectively) has a bit more power than the more unbalanced case (with 7 and 11 items), but the difference is very small.

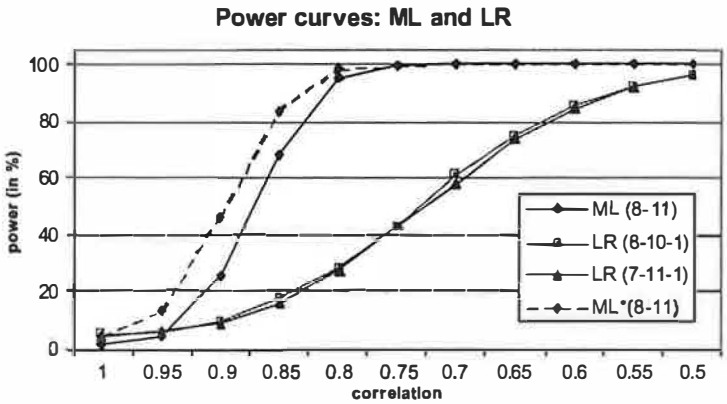

Figure 3.

As can be seen from Table 4, the splitter-item-technique has less power that the LR-test using three items for the criterion. The most powerful strategy is found when the three-item criterion is used to construct two contrasting groups (the cases 3 and 5). These two cases are compared with ML-test (nominal level) in Figure 4.

Here we see that the balanced case (8 items for each dimension) has more power than the unbalanced case (5 against 11 items), but the difference is very small compared to the difference in power with the ML-test.

In Figure 5 the power curves are given for the three LR-tests which yield the

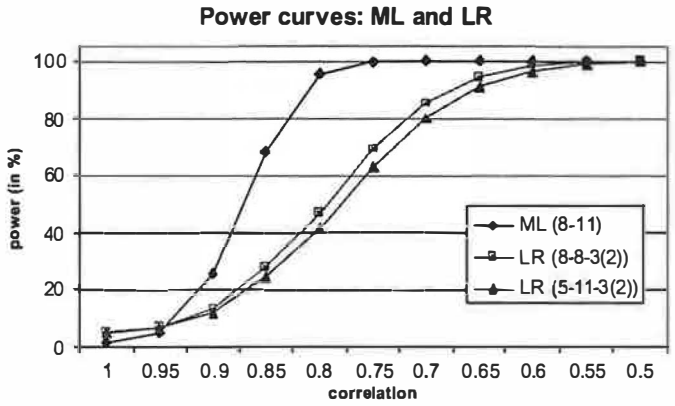

Figure 4

most balanced test halves (cases 1, 2 and 4; the notation in the legend, e.g., LR(8-8-3(2)) denotes the LR-test with 8 items in each dimension, 3 items used as criterion and the criterion test is used to form 2 contrasting groups). There it is seen that three items used for the criterion yields more power than using just one, and that using two contrasting groups yields more power than using four. A similar result was found for the three unbalanced cases (not displayed).

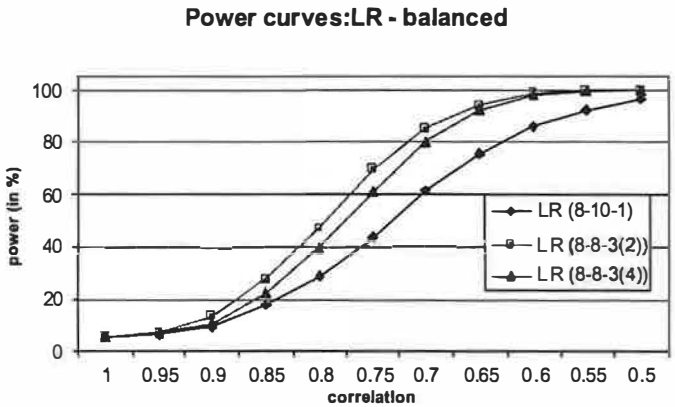

Figure 5.

It appears that balanced tests have more power than unbalanced ones, that a more extensive (and therefore more reliable) criterion yields more power and that a

contrast of two groups yields more power than a contrast of four groups.

If the criterion is made still more reliable, by making it longer, we might expect still more power, but with a given number of items, there will be a trade-off: making the criterion longer will make one test halve shorter (and the test more unbalanced). To explore this trade-off, two extra power curves were constructed with 5 and 7 items respectively in the criterion, and constructing two contrasting groups. The results are displayed in Figure 6. Most power is obtained with a criterion of 5 items; with 7 items the power is a little bit smaller. From this analysis we might conjecture that the LR-test has maximal power using a criterion of about half of the items loading on a single dimension and two contrasting groups. Of course this is a crude indication, which should be corroborated by more results, especially concentrating on differences in difficulty.

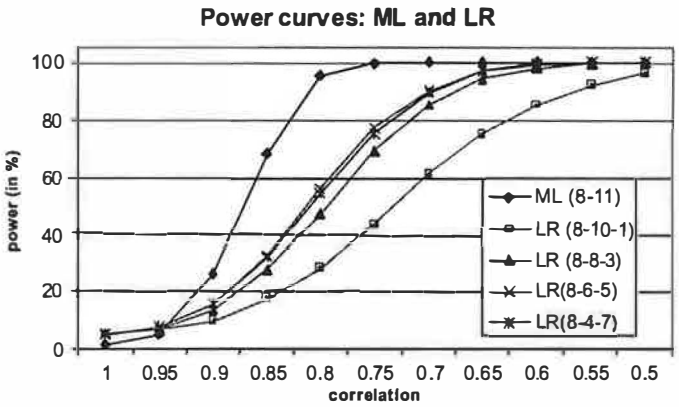

Figure 6

As a general result we can say that in the study reported here, the ML-test has considerably more power than the LR-test (the generalization of the splitter-item-technique). But this does not mean that it is always preferable to use the ML-test. In the simulation study it has been assumed that the partition of the item set in

two homogeneous subsets does not contain errors, an assumption which may not be true in reality, where an incorrect classification of items according to their relevant dimension may be rather the rule than the exception. The LR-test is in some sense less vulnerable to such an incorrect classification, because not all items have to be classified as belonging to one or the other dimension. The only way in which an incorrect classification may diminish power is in the selection of the criterion items: if the criterion consists of more than one item, a non-homogeneous criterion will affect the power. On the other hand, it is conjectured that an error in the composition of a multi-item criterion will have more dramatic consequences because the criterion consists of fewer items than the remaining test itself.

Before sorting this out, however, two problems of a more pure statistical nature have to be treated. The first one is the null distribution of the ML-test. Even with as few as 19 items, the null distribution is definitely not chi-square unless the sample size is huge. We did not succeed in deriving the true null distribution, and the fact that in the present example this distribution seems to be shrunken towards lower values (and thus delivers a conservative test) cannot simply be generalized to other cases.

The second problem has to do with the lack of power of the ML-test in case the correlation between the dimensions is high (which in practice will be more often the case than not).

The two problems are closely connected. As was shown earlier, the ML-test essentially compares the predicted and observed bivariate score table, and the longer the two test halves are, the more cells the table contains and at the same time the more degrees of freedom are associated with the test statistic. So the solution might be found in reducing the number of cells in some way and at the same time reducing

the number of degrees of freedom. How this can be accomplished will be explained in the next section.

A class of one-degree-of-freedom tests

Rationale

Under the null hypothesis we may consider the two test halves as representing each a latent variable, where the two latent variables correlate 1 with each other. As a consequence the frequencies in the bivariate score table will tend to be concentrated in a number of cells which we might label as the 'main diagonal'. Further down we will have to discuss what we exactly mean by this term, but for the time being we may consider the simple case of two test halves having an equal number of items, with similar distributions of the difficulty parameters, such that, in this case, the fuzzy term 'main diagonal' refers to the main diagonal of the bivariate frequency table. But even in such an ideal case, the correlation between the scores will be less than one, because of attenuation by unreliability, and therefore the observed frequencies of the off-diagonal cells will differ from zero.

If the correlation between the latent variables is less than one, the observed frequencies on the 'main diagonal' will tend to be less than in the unidimensional case, and the opposite effect will appear in cells farther away from the 'main diagonal'. So we can roughly split the cells of the bivariate table in a subset where, under the alternative model, the expected frequencies are larger than the observed ones, and the complementary subset where they will be smaller. The first subset will contain mainly cells along the 'main diagonal'. If we were successful in identifying these two subsets without inspecting the data, we might just sum the expected frequencies in one subset and compare this sum with the sum of the observed frequencies in the

same subset of cells. In other words, the tests will consist in taking 'cells together' such that the summed differences between observed and expected frequencies do not diminish due to cancellation.

Since there are two complementary subsets, we may expect to have a single degree of freedom.

The theory

In the ML-test the measurement model is the conditional Rasch model, and the marginal model is a saturated multinomial model for the score distribution. Both models are an exponential family and their combination also is an exponential family. Without loss of generality we can apply the same model also if all observations yielding a zero score or a perfect score are eliminated: for the conditional measurement model, these response patterns do not carry any information, and for the structural model they are simply eliminated from the sample space. So under this assumption, $k-1$ item parameters and $k-2$ multinomial parameters are to be estimated. The total number of different response patterns is $M = 2^k - 2$.

Now construct two matrices T_1 and T_2 of order $M \times k$ and $M \times (k-2)$ respectively, where each row corresponds to a response pattern. The row in T_1 is the response pattern itself, the row in T_2 is an indicator vector of the score: any two response patterns having the same score on the total test should have an identical row in T_2, containing a single 1-entry and zeros elsewhere.

Define the (partitioned) matrix U as

$$U = [T_1|T_2|Y], \tag{15}$$

where Y is an arbitrary matrix of constants with s columns. The matrix Y is the

most important part of the matrix U, and its construction usually needs a lot of creativity from the designer of the test, as will be discussed in the sequel. The matrix $[T_1|T_2]$ is added for technical reasons, namely to ensure that a certain function of U is asymptotically chi-squared distributed, no matter how Y is defined. This function is the quadratic form $Q(U)$, defined as

$$Q(U) = n(\mathbf{p}-\widehat{\boldsymbol{\pi}})'U[U'D_{\widehat{\pi}}U]^{-}U'(\mathbf{p}-\widehat{\boldsymbol{\pi}}), \tag{16}$$

where $\mathbf{p}$ is the M-vector of observed proportions of the response patterns, $\widehat{\boldsymbol{\pi}}$ is the M-vector of estimated probabilities, $D_{\widehat{\pi}}$ is a diagonal matrix with $\widehat{\boldsymbol{\pi}}$ as its main diagonal and the superscript '$-$' denotes a generalized inverse. The main and quite powerful result (Verhelst and Eggen, 1989; Glas and Verhelst, 1995 and Verhelst and Glas, 1995) can be formulated as follows.

1. If $\widehat{\boldsymbol{\pi}}$ is estimated using a BAN-estimator, then $Q(U)$ is asymptotically chi-square distributed. The associated degrees of freedom are given by

$$df(Q(U)) = \text{rank}(U) - \text{number of parameters } - 1 \tag{17}$$

2. If $\widehat{\boldsymbol{\pi}}$ is estimated using maximum likelihood, then

$$Q([T_1|T_2]) \equiv 0 \tag{18}$$

If we take care that the columns of Y are mutually linearly independent and independent of the columns of T_1 and T_2 as well, then the number of degrees of freedom is simply the number of columns of Y, i.e., s.

To make things a bit more concrete, an example with $k = 4$ items is displayed in Table 5. The column Y will be explained in the next subsection. Notice that the two response patterns yielding a zero score and a perfect score are omitted.

Table 5. An example of a U-matrix

T_1				T_2			Y
0	0	0	1	1	0	0	0
0	0	1	0	1	0	0	0
0	1	0	0	1	0	0	0
1	0	0	0	1	0	0	0
0	0	1	1	0	1	0	0
0	1	0	1	0	1	0	2
0	1	1	0	0	1	0	2
1	0	0	1	0	1	0	2
1	0	1	0	0	1	0	2
1	1	0	0	0	1	0	0
0	1	1	1	0	0	1	0
1	0	1	1	0	0	1	0
1	1	0	1	0	0	1	1
1	1	1	0	0	0	1	1

The algebra

For the example used in Table 5, we used $I_1 = \{1, 2\}$ and $I_2 = \{3, 4\}$. To construct the test, one needs a definition of what was described loosely as the main diagonal, but formally this is a subset of cells of the bivariate score table. We will denote this set as S:

$$S = \{(s_1, s_2) : \text{ cell of the 'main diagonal'}\}$$

For the example in Table 5, S has been defined as

$$S = \{(1, 1),\ (2, 1)\}$$

and as can be checked in the column Y of Table 5, all non-zero entries correspond to a response pattern of one of these cells, and no zero entry does belong to either

of these cells.

Although the matrix Y is arbitrary, a constraint will be put on it to keep it practicable (since the number of rows grows exponentially with k). Therefore a practical rule is introduced: all rows in Y corresponding to the same cell of the bivariate score table are identical. In Table 5, all entries fitting in cell $(1,1)$ have a weight equal to 2, and the rows fitting in cell $(2,1)$ have a weight of 1. Using different weights for different cells may reflect the user's certainty about the sign of the difference between expected and observed frequencies under the alternative hypothesis. Weights certainly will have an impact on the power of the test. Weights will be denoted by the symbols $v(s_1, s_2)$ or v_{s_1,s_2}.

Computing $Q(U)$ using (16) as a computational formula for the example in Table 5 is not too hard, but it is readily seen that problems will arise for larger k. For $k = 20$, the matrix has more than one million rows, and for every additional item, the number of rows doubles.

As a first step to simplify the computations, we will construct a matrix U of full rank (such that all generalized inverses are regular inverses). It is not difficult to check that the partitioned matrix $[T_1|T_2]$ in Table 5 is not of full rank; in fact its rank is one less than the number of columns. To get rid of this dependency we may discard one of the columns of T_1, the last one, say. This reduced matrix will be denoted T_1 as well. If $T = [T_1|T_2]$ is of full rank and the one column matrix Y is linearly independent of it, then, because of (18), the quadratic form $Q(U)$ can be rewritten as

$$Q(U) = n(\mathbf{p}-\widehat{\boldsymbol{\pi}})'Y[Y'D_{\widehat{\pi}}Y - Y'D_{\widehat{\pi}}T(T'D_{\widehat{\pi}}T)^{-1}T'D_{\widehat{\pi}}Y]^{-1}Y'(\mathbf{p}-\widehat{\boldsymbol{\pi}}), \qquad (19)$$

where

$$T'D_{\widehat{\pi}}T = \left[\begin{array}{c|c} T_1'D_{\widehat{\pi}}T_1 & T_1'D_{\widehat{\pi}}T_2 \\ \hline T_2'D_{\widehat{\pi}}T_1 & T_2'D_{\widehat{\pi}}T_2 \end{array}\right] = \left[\begin{array}{c|c} T_{11} & T_{12} \\ \hline T_{21} & T_{22} \end{array}\right], \tag{20a}$$

and

$$Y'D_{\widehat{\pi}}T = \left[\begin{array}{c|c} Y'D_{\widehat{\pi}}T_1 & Y'D_{\widehat{\pi}}T_2 \end{array}\right] = [E'|F'] . \tag{21}$$

Before deriving the elements of the needed matrices and vectors in detail, it may be useful to consider (19) in some other form. For the case the non-zero elements of the one-column matrix Y correspond to response patterns which belong to the subset S, (19) can be rewritten as

$$Q(U) = \frac{n}{K}\left[\sum_{(s_1,s_2)\in S} v_{s_1s_2}(p_{s_1s_2} - \widehat{\pi}_{s_1s_2})\right]^2 \tag{22}$$

where K is a 1×1 matrix: it is the expression between brackets in the right-hand side of (19). As is easily seen, the proposed test statistic is proportional to the square of the weighted sum of the differences between observed and expected cell proportions in the cells belonging to the set S. The art in finding a powerful test will consist in finding a definition of S, having as many elements as possible and such that all or most of these differences have the same algebraic sign under the alternative hypothesis. The computational difficulty is related to the evaluation of the quantity K, to which we return now.

The elements of matrices T_{11}, T_{12} and T_{22} are always present in tests of the general form (19) and are derived elsewhere (Verhelst and Eggen, 1989, p. 61; Verhelst and Glas, 1995, p. 227). They are repeated here for convenience. It will be assumed throughout that the rank number of the columns of T_2 equals the corresponding

score; this is the case in Table 5. In the formulae to follow, the subscripts i and j refer to items and the subscripts s and t refer to scores.

$$(T_{11})_{ij} = (T_1' D_{\widehat{\pi}} T_1)_{ij} = \frac{1}{n} \sum_{s=1}^{k-1} n_s \widehat{\pi}_{ij|s}, \quad (i,j = 1, \ldots, k-1), \tag{23}$$

where $\pi_{ij|s}$ is the conditional probability of having a correct response on items i and j given the score s, and where we define

$$\pi_{ii|s} = \pi_{i|s}.$$

Notice that n denotes the total number of response patterns not leading to a zero or maximum score.

Next,

$$(T_{22})_{st} = (T_2' D_{\widehat{\pi}} T_2)_{st} = \begin{cases} \frac{n_s}{n} \text{ if } s = t, \\ \\ 0 \text{ otherwise.} \end{cases}, \quad (s,t = 1, \ldots, k-1), \tag{24}$$

and

$$(T_{12})_{is} = (T_1' D_{\widehat{\pi}} T_2)_{is} = \frac{n_s}{n} \widehat{\pi}_{i|s}, \quad (i,s = 1, \ldots, k-1) \tag{25}$$

For the other expressions, we collect all item parameters ε_i, $i \in I_\ell$ in a vector $\boldsymbol{\varepsilon}_\ell$ $(\ell = 1, 2)$ in an arbitrary but fixed order. Using Table 5 as an example it is not too hard to check the following expressions:

$$Y' D_{\widehat{\pi}} Y = \sum_{(s_1,s_2) \in S} \frac{n_{s_1+s_2}}{n} \frac{\gamma_{s_1}(\widehat{\boldsymbol{\varepsilon}}_1) \gamma_{s_2}(\widehat{\boldsymbol{\varepsilon}}_2)}{\gamma_{s_1+s_2}(\widehat{\boldsymbol{\varepsilon}}_1, \widehat{\boldsymbol{\varepsilon}}_2)} v^2_{s_1,s_2}, \tag{26}$$

$$(E')_i = (Y' D_{\hat{\pi}} T_1)_i = \sum_{(s_1,s_2)\in S} \frac{n_{s_1+s_2}}{n} \frac{\hat{\varepsilon}_i \gamma^{(i)}_{s_1-1}(\hat{\varepsilon}_1)\gamma_{s_2}(\hat{\varepsilon}_2)}{\gamma_{s_1+s_2}(\hat{\varepsilon}_1,\hat{\varepsilon}_2)} v_{s_1,s_2}, \quad (i \in I_1), \tag{27}$$

and of course, a similar expression holds if $i \in I_2$. And finally

$$(F')_s = (Y' D_{\hat{\pi}} T_2)_s = \sum_{\substack{(s_1,s_2)\in S \\ s_1+s_2=s}} \frac{n_{s_1+s_2}}{n} \frac{\gamma_{s_1}(\hat{\varepsilon}_1)\gamma_{s_2}(\hat{\varepsilon}_2)}{\gamma_{s_1+s_2}(\hat{\varepsilon}_1,\hat{\varepsilon}_2)} v_{s_1,s_2}, \quad (s = 1,\ldots,k-1). \tag{28}$$

Using the rule for an inverse of a partitioned matrix, it follows that

$$[T' D_{\hat{\pi}} T]^{-1} = \left[\begin{array}{c|c} B & C \\ \hline C' & D+\Delta \end{array}\right], \tag{29}$$

where

$$B = \left[T_{11} - T_{12}T_{22}^{-1}T_{21}\right]^{-1}, \tag{30}$$

$$C = -BT_{12}T_{22}^{-1}, \tag{31}$$

$$D = T_{22}^{-1}T_{21}BT_{12}T_{22}^{-1} = -T_{22}^{-1}T_{21}C, \tag{32}$$

and

$$\Delta = T_{22}^{-1}. \tag{33}$$

Verhelst and Eggen (1989, p. 62) show that the elements of the matrix B^{-1} are equal to the elements of minus the matrix of second derivatives of the conditional loglikelihood, divided by the effective sample size n. In a conditional estimation procedure, the matrix B^{-1} is usually computed to yield estimates of the standard

errors of the item parameter estimates.

Defining

$$G = T_{12}T_{22}^{-1}, \text{ i.e., } (G)_{is} = \widehat{\pi}_{i|s}$$

and using (30), (31)and (32) it follows that

$$Y'D_{\widehat{\pi}}T(T'D_{\widehat{\pi}}T)^{-1}T'D_{\widehat{\pi}}Y = (E - GF)'B(E - GF) + F'\Delta F \tag{34}$$

which, subtracted from (26) gives the 1×1 matrix K, to be used in (22).

Defining the 'main diagonal'

Since the rationale of the test consists in 'taking cells together', the power of the test will critically depend on the definition of the set S. If this set is defined in such a way that the difference between observed and expected frequencies has the same sign for all the cells in the set, then no cancellation will occur. Moreover, it may be argued that the set S should be as large as possible. But of course, the deviations themselves must not be inspected to define S.

If $k_1 = k_2$ and both test halves are equally difficult, then the main diagonal of the bivariate frequency table may be a suitable choice for S, but in other cases the choice is not obvious. So, one has to develop one or more reasonable heuristics to define S. We developed two heuristics which are based on the following rationale. If the item parameters are given (under the null hypothesis), then one can rank the cells with the same total score according to their expected frequencies, and assign one or more cells with high expected frequencies to the set S. For the small power study we undertook, the following two heuristics have been applied:

1. For each score $2 \leq s \leq k-2$, the cell with the highest expected frequency belongs

to S.

2. As heuristic 1, but for each score $5 \leq s \leq k-5$ and s odd, the cell with the next highest frequency also belongs to S.

For the same test specifications as in earlier studies ($k_1 = 8$, $k_2 = 11$, all item parameters equal zero, $n = 1000$ and number of replications is 4000), the power of the two one-degree-of-freedom tests has been estimated. The results, together with the Martin-Löf test and the splitter-item-test (taking one item from the largest test halve as criterion) are given in Table 6 and are graphically displayed in Figure 7. The differences are most dramatically seen for $\rho = 0.85$ where the power ranges from less than 20% for the splitter-item-technique to more than 90% for the one degree of freedom test using heuristic 2.

Table 6. Comparison of power

ρ	ML	LR	$X^2(1)$	$X^2(2)$
1.00	1.55	5.10	5.27	5.28
0.95	5.02	6.12	15.9	21.8
0.90	25.8	9.50	48.7	62.4
0.85	68.2	17.4	78.7	90.9
0.80	95.2	28.4	93.6	98.9
0.75	99.7	43.2	98.7	99.9
0.70	100	61.5	99.9	100
0.65	100	75.3	100	100
0.60	100	85.6	100	100
0.55	100	92.3	100	100
0.50	100	96.6	100	100

Notice that there is no simple relationship between the power of the tests and their associated number of degrees of freedom: the ML-test has 87 degrees of freedom, the splitter-item-technique (with the least power) has 17, while the two newly

introduced tests, with the greatest power, have one degree of freedom. As to the (considerable) difference in power between the latter two, it may be noticed that the average (across replications and values of ρ) percentage of the summed expected frequencies of the cells belonging to S relative to the effective sample size is 36% for heuristic one, and 47% for heuristic two, suggesting that the power increases the more the percentages of observations covered by the set S and its complement approach 50%. Of course, more than the evidence presented here is needed to rely on this suggestion.

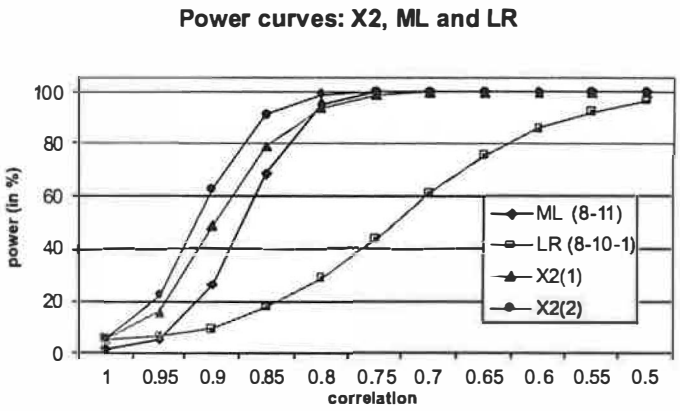

Figure 7

What if one is mistaken?

The Martin-Löf test and the one-degree-of-freedom tests developed here have a quite specific alternative hypothesis: the partition of the items along the two dimensions is highly specific, and with a considerable number of items in the test it is quite likely that some errors may occur in the partitioning with respect to the two dimensions. In the splitter-item-technique (using a single item as a criterion) it does not matter very much to which dimension the criterion item belongs, as long as the test is reasonably balanced with respect to the two dimensions. (Compare

the cases 1 and 2 in Table 4.) In the generalization of the splitter-item-technique, taking more than one item as a criterion, the consequences of a specification error may be dramatic for the power of the test, as can be seen if two equally difficult items are taken as criterion: no matter how the high and low groups are defined, the two dimensions will be equally represented in both groups and the test will loose all its power.

In the ML-test and the one-degree-of-freedom tests, things may be less dramatic as long as the majority of the items in each test halve represent the same dimension. To get an idea about the loss of power two extra power studies have been undertaken, using only heuristic 2 of the one-degree-of-freedom tests. In the first study two items are incorrectly classified and in the second study four erroneous classifications occur, but in both studies, the two test halves have 8 and 11 items respectively. The results are displayed graphically in Figure 8.

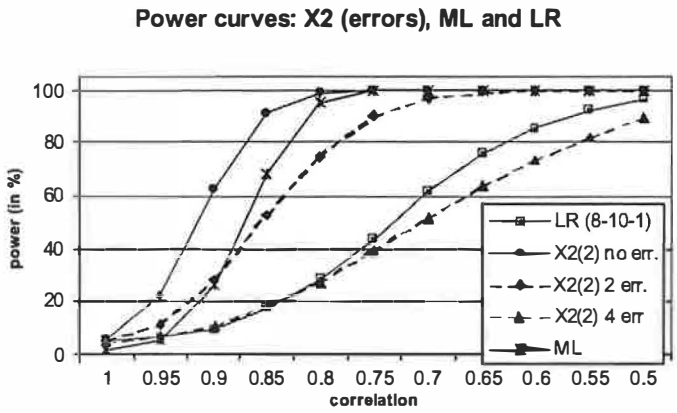

Figure 8

The power drops quite dramatically if errors in the assignment to the dimensions are made. If four of the 19 items are incorrectly classified, the test performs worse than the splitter-item-technique. If two errors are made, the power is higher than the

power for the ML-test if the correlation is higher than 0.90, but as the correlation decreases, the power raises less steeply than the power of the ML-test. Although the power curve in the case of four specification errors may be disappointing, it should be realized that the first dimension is identified by only 8 items, two of which are erroneously assigned to this dimension, which may be considered as a rather poor identification of the dimensionality of the total item set.

To get an impression of the power in a more realistic case, the power curves have been determined also for the case where $k_1 = k_2 = 15$, for identifications with zero, two, four and six errors respectively. In the worst case, the subset identifying each dimension contains 12 items of that dimension and three items of the other dimension. The results are displayed graphically in Figure 9.

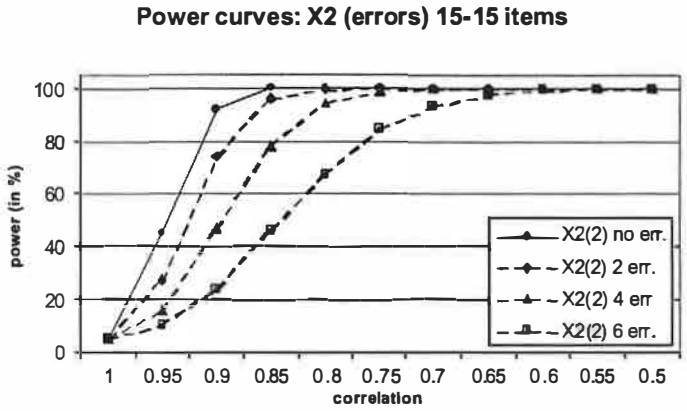

Figure 9

It can be seen from this Figure that the power drops rather quickly (yielding a power of less than 50% in case of 6 errors and a correlation of 0.85). On the other hand, the power curves with 30 items are much steeper than in the case with 19 items. In the errorless case and with a correlation of 0.95, the power with 19 items is just above 20%, while with 30 items it is 45%.

One may regret this rather drastic loss of power in the case of specification errors, or even find the requirement of a correct identification of the dimensions too stringent for the proposed tests to be useful in practice, and favor an alternative approach which does not require any a priori identification of the dimensions. The splitter-item -technique is such a procedure, but the above results have shown that its power is poor in comparison to the Martin-Löf test or the one-degree-of-freedom tests introduced here. There are other approaches possible, and one of them will be discussed in the next section.

Other possibilities

It may be interesting to see how versatile the general theory about test construction is, on the one hand, and on the other hand to discuss its limitations. The matrix Y in Table 5 has 1 column, but other matrices with more columns (yielding tests with more degrees of freedom) may be constructed as well. We will consider three examples in the domain of testing unidimensionality: the X^2-analogue of the Martin-Löf test, the R_{2c}-test developed by Glas (1989), and a specialization of this test, which takes advantage of prior information about the dimensionality.

1. the X^2-analogue of the Martin-Löf test. It has been shown in a previous section (see equation (11)) that the ML-test statistic can be decomposed as the sum of two likelihood ratio statistics, one of which is very dominating and amounts to a likelihood ratio statistic comparing the expected bivariate frequency table to the observed table. But such a comparison could of course be done via a X^2-like statistic, and the precise form is given by the quadratic form $Q(U)$ of equation (16), with a Y-matrix having a column for each cell of the bivariate frequency table (with the exception of the $(0,0)$ and the (k_1,k_2) cells, which

do not contribute to the test statistic). The entries corresponding to response patterns which belong to this cell equal one, the others are zero.

2. In 1989 Glas, following a rationale put forward by van den Wollenberg (1979,1982), developed the R_{c2}-test which essentially compares the expected and observed frequencies in the 2×2 tables formed by all item pairs. Although Glas followed another rationale than the one presented above, it can be shown that this test can also be considered as a quadratic form $Q(U)$, where the Y-matrix has a column for each item pair, with a 1 entry for all response patterns having a correct answer to both items of the pair. The number of degrees of freedom in this case equals the number of pairs of items.
3. If item parameters are estimated using a unidimensional conditional Rasch model and a saturated multinomial model for the distribution of the test scores, it seems plausible that the predicted 2×2 frequency table will show a closer correspondence to the observed table for pairs of items where both items belong to the same dimension (homogeneous pairs), than for pairs where the items belong to different dimensions (heterogeneous pairs). Consequently, more power may be achieved if the matrix Y is restricted to the heterogeneous pairs, instead of using a column for each pair as in the R_{2c}-test. The number of columns in this case is then $k_1 k_2$, while in the R_{2c}-test $k(k-1)/2$ columns are required.

The problems associated with the approach where Y has many columns, however, are not always easy to solve. We discuss some of the problems occurring when one tries to construct tests like the ones in the examples above.

1. The rank of the U-matrix has to be determined on theoretical grounds. For quite general approaches this is usually not trivial. For the U-matrix in the first ex-

ample, it is not hard to show that it is not of full rank, but is much harder to show what its rank is in general.

2. The computational burden is substantial if the matrix $Y'D_{\hat{\pi}}Y$ is not diagonal (which it is not in each of the three examples above), i.e., when the columns of Y are not mutually orthogonal. For the R_{2c}-test, for example, basic symmetric functions must be computed for the parameter vector with all singletons, pairs, triples and quadruples of parameters removed. If the number of items is substantial, computing time may become prohibitive.
3. But even if computing time is not a problem, numerical instability may become a big problem. For the R_{2c}-test, the matrix of the quadratic form $Q(U)$ is the inverse of a symmetric matrix with $k(k-1)/2$ rows and columns. It is quite a complicated problem to show the numerical accuracy of the computed result of $Q(U)$, even for a moderate number of items (say 30 to 40).
4. Reliance on asymptotic results may become problematic. As was shown in Figure 2, the null distribution of the ML-statistic differs substantially from the asymptotic chi square distribution (with 1000 observations, the real significance level is 1.5% where the nominal level is 5%). For the X^2-analogue of this test the situation is not different: the expected bivariate frequency table is used here as well. Similar problems may appear with the R_{2c}-test: the expected proportion of people having two correct answers can be extremely low (or high) for pairs of items which are both difficult (or easy).

None of the above problems occurs when a one-degree-of-freedom test is constructed. With a one-column matrix Y the number of degrees of freedom is either one or zero (and the latter case occurs only if the Y-vector lies in the column space of $[T_1|T_2]$, which is usually not too hard to check), the computational burden is

relatively low, the matrix to be inverted is trivially simple (1×1), and asymptotic results will quickly apply if, as in heuristic 2, the summed expected frequencies do not differ too much from half of the effective sample size. Moreover, the total number of items in the test does not complicate the computations in any important way, while the computation of the R_{2c}-statistic gives problems if $k > 15$ (Glas, 1989, p. 42 and p. 97).

Conclusion

A number of statistical tests especially designed to test the unidimensionality axiom of the Rasch model have been investigated and compared with respect to their statistical power. It may be interesting to look at these tests from the viewpoint of the risk taken by the user and the pay-off for taking and not taking risk.

If the user does not take any risk, leaving the testing as a mechanical routine to his software, he can use the splitter-item-technique (and, for example, use each item in turn as a splitter item) to detect violation of the unidimensionality axiom or he could use an omnibus test like the R_{2c} (if available). For the splitter-item-technique, it has been shown that the power is not very high, which will probably also contribute to the finding that when the items are used in turn as a criterion, a proportion of the tests will lead to significant results and the others not, without showing a clear-cut or easy to interpret pattern.

On the other hand, if the user has a priori information on the dimensional structure, the power can be increased substantially. Three different procedures may be followed here, and their pro's and con's can be summarized as follows.

1. A generalization of the splitter-item-technique where the criterion consists of the score on a homogeneous subtest. From the power studies reported here, it appears

that the power increases substantially as compared to the splitter-item-technique (see Figure 5), and that the maximum power is reached when approximately half of the items of one dimension are used for the criterion. It seems also that the test has more power if the sample is split into two contrasting groups rather than more than two. The test is a likelihood ratio test, and the number of degrees of freedom is

$$df = (G - 1)(k - k_c - 1)$$

where G is the number of contrasting groups, and k_c is the number of items used as a criterion. Although no detailed study has been undertaken on the loss of power if specification errors are made in the composition of the criterion, yielding a heterogeneous criterion, it is easy to see that there is no power if $k_c = 2$ and the criterion is not homogeneous.

2. The Martin-Löf test of unidimensionality, which essentially compares the bivariate frequency table with the expected one appears to have substantially more power than the optimal LR-test of the previous class (see Figure 5). Application of the test is relatively easy, but it requires full identification of the dimensions. The problem with this test is that the null distribution deviates substantially from the asymptotic distribution, unless the sample size is huge. The comforting finding of the present study is that the test seems to be conservative, and that one might therefore raise the nominal significance level to have a better approximation to the intended level, or conversely that the real power is greater than the power suggested by the estimated power curves. One should, however, be careful with this statement, because there is no theoretical evidence that the test is indeed conservative. A more accurate approximation of the null distribution

(in samples of realistic size) than the asymptotic one would be highly welcomed.

3. A class of generalized X^2-tests with one degree of freedom. The rationale of these tests consists in summing the deviations between observed and expected frequencies in the bivariate frequency table for a set of cells (the set S) where the algebraic sign of the deviations is expected to be the same. The power of the test will largely depend upon the accuracy with which this set can be identified without inspecting the deviations themselves. In the present study, two heuristics have been applied, which both yielded tests with substantially more power than the ML-test. For the best of the two heuristics the influence of specification errors on the power curves has been investigated. The decrease in power seems to be substantial in the case where for each dimension 12 of the 15 items belong to the same dimension (see Figure 9). Although not investigated in detail, it seems reasonable to expect similar effects for the other heuristic as well as for the ML-test.

As a concluding note, some suggestions for future research can be made. It has been shown that the general theory on constructing X^2-tests can deliver very powerful tests. The theory itself serves as a general framework, and the power obtained depends largely on the creativity of the user to define an adequate Y-matrix. The two heuristics presented above are perhaps not the most powerful ones, and endless variation (also using different weights) and good ideas may be necessary to build superior statistical tests with power against specific violations of the model used. In testing the unidimensionality of the Rasch model, one might think, for example, of assigning negative weights to the complement of S, or to form three sets of cells, receiving positive, negative and zero weights, respectively. From an application point of view, the availability of software where the user can define his own tests by spec-

ifying the Y-matrix, would be highly welcomed as a powerful and original research tool.

References

Glas, C.A.W. (1989). *Contributions to estimating and testing Rasch models.* (Doctoral thesis.) Enschede: University of Twente

Glas, C.A.W. and Verhelst N.D. (1995). Testing the Rasch model. In G.H. Fischer and I.W. Molenaar (Eds.), *Rasch models: foundations, recent developments and applications* (pp. 69-95). New York: Springer-Verlag.

Gustafsson, J.E. (1980). Testing and obtaining fit of data to the Rasch Model. *British Journal of Mathematical and Statistical Psychology, 33,* 205-233.

Martin-Löf, P. (1973). *Statistiska modeller* [Statistical models.] Anteckningar från seminarier lasåret 1969-1970, utarbetade av Rolf Sundberg. Obetydligt ändrat nytryck, October 1973. Stockholm: Institütet för Försäkringsmatemetik och Matematisk Statistisk vid Stockholms Universitet.

Molenaar, I.W. (1983). Some improved diagnostics for failure of the Rasch model. *Psychometrika, 48,* 49-72.

OECD (2001). *Knowledge and Skills for Life, Rirst results from PISA 2000.* OECD: Paris

Van den Wollenberg, A.L. (1979). *The Rasch model and time limit tests.* (Doctoral thesis.) Nijmegen: University of Nijmegen.

Van den Wollenberg, A.L. (1982). Two new test statistics for the Rasch model. *Psychometrika, 47,* 123-139.

Verguts, T. and De Boeck, P. (2000). A note on the Martin-Löf test for unidimensionality. *Methods of Psychological Research Online, 5,* 77-82.

Verhelst, N.D. (1993). Item respons theorie. In T.J.H.M. Eggen and P.F. Sanders (Eds.), *Psychometrie in de praktijk* (pp. 83-178). Arnhem: Cito.

Verhelst N.D. and Glas, C.A.W. (1995). The one parameter logistic model. In

G.H. Fischer and I.W. Molenaar (Eds.), *Rasch models: foundations, recent developments and applications* (pp. 215-237). New York: Springer-Verlag.

Verhelst, N.D. and Eggen, T.J.H.M. (1989). *Psychometrische en statistische aspecten van peilingsonderzoek* [Psychometric and statistical aspects of assessment research.] (PPON-rapport , 4) Arnhem: Cito.