KM 5(1982)
pag 30-55

# ON THE APPLICABILITY OF THE $Q_2$ TEST FOR THE RASCH MODEL

Arnold L. van den Wollenberg*

Abstract

The $Q_2$ statistic as introduced by Van den Wollenberg (1979,1982) presumes a partitioning of the dataset into groups of subjects having the same raw score. Molenaar (1980) showed that the item parameters should be estimated within each subsample separately in order to obtain statistics, which are asymptotically distributed as chi-square.

In the present study it is shown that it is also possible to obtain the $Q_2$ statistic for composite partitionings, in which several level groups are combined into one subsample.

Simulation results are presented indicating that the $Q_2$ statistic for the alternative partitionings approximates the chi-square distribution to a satisfactory degree. Violation of the unidimensionality axiom is clearly detected in the partitionings.

It is argued that by these results the applicability of the $Q_2$ statistic is greatly improved.

* Department of Mathematical Psychology
University of Nijmegen
P.O.Box 9104
6500 HE Nijmegen

## 1. Introduction

In the last few years the Rasch model has rapidly gained popularity among social scientists. This should not at all be surprising, considering the desirable characteristics of the model deriving from the central property of specific objectivity. One of these characteristics of the Rasch model is sample independence (°) which says:

> When the model holds for a given population of subjects, every sample from that population, may it be random or not, will yield the same item parameter estimates within sampling fluctuations.

It is this required equality of item parameters over subsampless which has been used to device statistical tests for the model. The test of Fischer and Scheiblechner (1970) inspects the equality of the item parameters of two disjunct subsamples explicitely by means of the statistic:

$$(1) \qquad S_i = \frac{\hat{\sigma}_i^{(1)} - \hat{\sigma}_i^{(2)}}{\{s^2_{\hat{\sigma}_i^{(1)}} + s^2_{\hat{\sigma}_i^{(2)}}\}^{\frac{1}{2}}} \quad .$$

Here $\hat{\sigma}_i^{(1)}$ is the parameter estimate of item i in the first subsample and $s_{\hat{\sigma}_i^{(1)}}$ is the corresponding standard error of estimate. For sufficiently large N the squared statistics (1) are distributed as chi-square with one degree of freedom; summation of the statistics (1) over all items then gives an overall statistic, which is, according to Fischer and Scheiblechner, distributed as chi-square with k-1 degrees of freedom. Van den Wollenberg (1979) showed that this last contention is not

---

(°) For a detailed discussion of sample independence and specific objectivity one is referred to Fischer (1974).

true. For our present purposes it suffices to point out that the statistic checks on the equality of the item parameters of the subsamples in the partitioning. The number of subsamples in the Fischer-Scheiblechner procedure is always two.

The conditional likelihood ratio test (Andersen, 1973) inspects whether the overall conditional likelihood is equal to the product of the likelihoods of the subsamples in the partitioning:

$$(2) \qquad \lambda = \frac{L\ (\hat{\underline{\varepsilon}})}{\prod_r L_r\ (\hat{\underline{\varepsilon}}^{(r)})} \simeq 1 \quad .$$

Here $L\ (\hat{\underline{\varepsilon}})$ is the maximum of the conditional likelihood function in the total sample, given the item parameter estimates, $\hat{\underline{\varepsilon}}$, of the total sample and $L\ (\hat{\underline{\varepsilon}}^{(r)})$ is the maximum of the likelihood function in the subsample with raw score r, given the item parameter estimates of that subsample. The approximate equality (2) becomes an exact equality, when the item parameters of all subsamples are equal and, the other way around, the equality (2) can only hold when the item parameters are equal. So the conditional likelihood ratio test checks the equality of the item parameters and this equality alone.

The statistical tests of Martin Lof (1973), Wright and Panchapakesan (1969) and the $Q_1$ statistic of Van den Wollenberg (1979,1982) all entail a comparison of expected and observed frequencies in the subsamples, given the item parameter estimates of the total sample. The situation will be depicted for the $Q_1$ statistic; the observations made in the following with respect to $Q_1$, hold to the same extent for the other two statistics.

Under the null-hypothesis that the total sample item parameters are equal to the item parameters of level group r, the probability that item i will be solved by a subject with raw score r is:

$$(3) \qquad \hat{\pi}_{ri} = \frac{\hat{\varepsilon}_i \, \hat{\gamma}^{(i)}_{r-1} \, (\underline{\hat{\varepsilon}})}{\hat{\gamma}_r \, (\underline{\hat{\varepsilon}})} \quad .$$

Here $\hat{\gamma}_r$ is the elementary symmetric function and $\hat{\gamma}^{(i)}_{r-1}$ its partial derivative with respect to item parameter $\hat{\varepsilon}_i$ (see e.g. Fischer (1974), Van den Wollenberg (1979)). The expected frequency may now be obtained by multiplying (3) by the number of subjects in level group r:

$$(4) \qquad E(n_{ri}) = n_r * \frac{\hat{\varepsilon}_i \, \hat{\gamma}^{(i)}_{r-1}}{\hat{\gamma}_r} \quad .$$

This expected frequency under the model is compared with the observed frequency by means of the familiar Pearson statistic:

$$(5) \qquad q_{ri} = \frac{\{n_{ri} - E(n_{ri})\}^2}{E(n_{ri})} + \frac{\{n_{r\bar{i}} - E(n_{r\bar{i}})\}^2}{E(n_{r\bar{i}})} \quad ,$$

where $n_{r\bar{i}}$ stands for the number of subjects in level group r with a negative response on item i. For further details concerning the rational of the statistic one is referred to Van den Wollenberg (1982).

When the item parameters of level group r would be estimated, which in fact is not the case, the estimation equation would be:

$$(6) \qquad n_{ri} = n_r * \frac{\hat{\varepsilon}^{(r)}_i \, \hat{\gamma}^{(i)}_{r-1} \, (\underline{\hat{\varepsilon}}^{(r)})}{\hat{\gamma}_r \, (\underline{\hat{\varepsilon}}^{(r)})} \quad .$$

The difference between (4) and (6) is only lying in the item parameters, so again the conclusion must be that the statistic at hand implies a comparison of the equality of the item parameters over subsamples, which holds in the same way for the Martin Lof statistic. The Wright-Panchapakesan statistic was shown to be in error (Van den Wollenberg (1980)), but when the errors are corrected the statistic becomes equivalent to $Q_1$ and by that the above observations also hold for this statistic.

The statistics of Martin Lof (1973), Wright and Panchapakesan and the $Q_1$ statistic of Van den Wollenberg (1979,1982) mount up to a check on the equality of item parameters over subsamples, even though the item parameters are only estimated for the total sample and not for the subsamples.

Several authors (Gustafsson and Lindblad (1978), Stelzl(1979) and Van den Wollenberg (1979,1982)) have pointed out that the above tests inspecting equality of item parameters, may fail to detect violation of the dimensionality axiom. Van den Wollenberg (1979) gave a set of sufficient conditions under which equality of item parameters over subsamples was realized, even for a two-dimensional latent space.

Equality of item parameters over subsamples may be a necessary condition for the Rasch model to hold, it most certainly is not a sufficient condition. The model tests of the above type are especially sensitive to violations of the axioms of monotonicity and sufficiency (Van den Wollenberg (1979), Gustafsson(1980a)); it can be said that these tests inspect the parallellism of item characteristic curves (Gustafsson (1980a)).

Van den Wollenberg (1979,1982) pointed out that the failure of these test procedures to detect certain types of model violations is associated to the partitioning of the sample according to raw score, which is the most commonly used partitioning. However, this is not to say that any other partitioning of the sample into subsamples will not suffer from this inadequacy. Van den Wollenberg (1979,1981) introduced a method to test the dimensionality axiom, using the above statistics, which is

based on a suitable partitioning of the dataset using items of the test as partitioning criterion. These items are called splitter items by Molenaar (1980).

Van den Wollenberg (1979,1982) also introduced a new statistic, $Q_2$, which is especially sensitive to violation of the axioms of unidimensionality and local stochastic independence. In this paper we will focus our attention on the practical applicability of this statistic. In section 2 this statistic will be introduced shortly. In section 3 some problems in the application of $Q_2$ will be discussed, whereas in section 4 a possible solution to the problems will be introduced. In this section some general observations will be made with respect to the application of $Q_2$ in practical settings.

## 2. The $Q_2$ statistic

The $Q_2$ statistic was introduced by Van den Wollenberg (1979,1982) in order to fill the existing gap in the testing equipment for the Rasch model. The statistic was shown to be especially sensitive to violation of the dimensionality axiom. In fact, the statistic inspects local stochastic independence of item pairs within level groups. However, when more dimensions underly the data, this will show in lack of local stochastic independence,when a one-dimensional model is applied. It is obvious that $Q_2$ is by nature also sensitive for violation of local stochastic independence, that does not derive from violation of unidimensionality.

For a detailed discusion of $Q_2$ one is referred to Van den Wollenberg (1982). Now we will make do with a short exposition of the procedure to obtain the statistic.

- The total sample is divided into level groups according to raw score. As usual the level groups 0 and k are excluded from the analysis leaving k-1 subsamples (1,.....,k-1).

- The level groups 1 and k-1 are also excluded from the analysis for obvious reasons. In the level group with raw score 1 the second order frequencies (i+,j+) are necessarily equal to zero (otherwise the raw score would at least be 2 in stead of 1) and quite analogously the second order frequencies (i-,j-) are zero in the level group with raw score k-1. So these level groups cannot give information about the association between the item pairs, and hence are removed from the sample. This implies that a total of k-3 (r=2,....,k-2) level groups is retained in the $Q_2$ analysis.

- For each level group item parameters are estimated and by means of these estimates the second order probabilities are obtained:

$$(7) \qquad \hat{\pi}_{rij} = \frac{\hat{\varepsilon}_i \, \hat{\varepsilon}_j \, \hat{\gamma}_{r-2}^{(i,j)}}{\hat{\gamma}_r} \quad ,$$

where $\hat{\pi}_{rij}$ is the estimated probability of a simultaneous realization of i and j and $\hat{\gamma}_{r-2}^{(i,j)}$ are the second order partial derivatives of the elementary symmetric functions with respect to $\hat{\varepsilon}_i$ and $\hat{\varepsilon}_j$ .

- Comparison of observed and expected second order frequencies of the 2*2 contingency table is again performed by means of the Pearson statistic:

$$(8) \qquad q_{rij} = \frac{D^2}{E(n_{rij})} + \frac{D^2}{E(n_{ri\bar{j}})} + \frac{D^2}{E(n_{r\bar{i}j})} + \frac{D^2}{E(n_{r\bar{i}\bar{j}})} \quad ,$$

where $D^2$ is the squared difference between observed and expected

frequency, which is equal for all cells in the 2*2 contingency table. Van den Wollenberg (1979,1982) claims that (8) is asymptotically distributed as chi-square with one degree of freedom. Molenaar (1980) studied four different ways to construct a $Q_2$ type statistic. The four cases are obtained by using either overall item parameter estimates or the estimates of each level group and by conditioning upon the marginals $n_r$, $n_{ri}$ and $n_{rj}$ or only upon $n_r$. He showed that a satisfactory statistic is only obtained when conditioning is performed on all three marginals mentioned above and when, simultaneously, level group parameter estimates are used. This is exactly the procedure followed above.

The overall statistic for level group r is obtained by summation over all item pairs:

$$(9) \qquad Q_{2r} = \frac{k-3}{k-1} \sum_i \sum_j q_{rij} \qquad \begin{matrix}(i = 1,\ldots,k-1)\\(j = i,\ldots,k\ )\end{matrix} \ ,$$

where (k-3)/(k-1) is a factor correcting for the covariance of the individual statistics $q_{rij}$.

It is assumed that the statistic $Q_{2r}$ is asymptotically distributed as chi-square with $\frac{1}{2}k(k-3)$ degrees of freedom. This in fact amounts to the assumption that the covariance between the individual $q_{rij}$ statistics can be accounted for by the correction factor (k-3)/(k-1). For the $Q_1$ statistic it has been shown (Van den Wollenberg, 1982) that a similar factor can be derived analytically, when all item parameters are equal. This proof is based on the fact that equal item parameters imply equal covariances. A similar proof should be possible for the present case, but has as yet not been provided. Simulation studies show that for the case of equal item parameters $Q_{2r}$ can at least be closely approximated by a chi-square distributed variate. When the item parameters are not

equal, the approximation is a bit worse, but still quite satisfactory (Van den Wollenberg (1982)).

The individual level group statistics $Q_{2r}$ can be summed to obtain an overall statistic $Q_2$ :

$$(10) \qquad Q_2 = \sum_r Q_{2r} \qquad (r = 2,\ldots,k-2) \quad .$$

As the $Q_{2r}$ (r=2,...k-2) are independent, $Q_2$ is chi-square distributed, when the $Q_{2r}$ statistics are. The addition of the $Q_{2r}$ statistics is yet not as straightforward as it may seem. Each $Q_{2r}$ statistic is obtained conditional upon the parameter estimates of the level group involved. So one could say that each $Q_{2r}$ is in fact testing a slightly different null-hypothesis. However, when $Q_2$ is used after a test of the $Q_1$ type has been performed, this does not seem to be serious, as in each case deviations from local independence are assessed.

Although $Q_2$ is defined as the statistic (10), we will also use the term in a generic sense to indicate the whole testing procedure.

## 3. Some difficulties in the application of $Q_2$

From the fact that the $Q_2$ analysis is performed on every level group separately, several difficulties arise.

1 Computing time. For every level group the item parameters have to be estimated, which can be a time-consuming affair, when large item pools and subject samples are involved.

2 Impossibility of estimation. When in a subsample an item has been passed or failed by all subjects, parameter estimation will prove impossible. This is especially likely to occur in the high and the low scoring subsamples: a difficult item will rarely be passed in a low score subsample, whereas an easy item will rarely be failed in a high score group of subjects. Deletion of items for this reason is very unattractive, because these items can be very relevant for other level groups.

3 Instability of statistic. When a $Q_2$ analysis is performed on a k item test, there will be k-3 relevant level groups. Within each level group the items are inspected pairwise, a total of k(k-1)/2 pairs. For each item pair a 2*2 contingency table is constructed, so all in all

$$4*k(k-1)/2*(k-3) = 2k(k-1)(k-3)$$

cells are involved in the observation matrix. The number of cells is a rapidly increasing function of k; for some values of k the number of cells is given here:

| k | cells |
|---|---|
| 5 | 80 |
| 10 | 1250 |
| 15 | 5040 |
| 20 | 12920 |
| 25 | 27600 |

Each subject figures in each 2*2 table, so the number of observations is equal to $\frac{1}{2}k(k-1)$. When for instance 5 observations per cell would be required, 20 subjects per level group would be a minimum and for a 25 item test at least 440 subjects are needed in the relevant

level groups.

This minimum value of N is not sufficient to prevent very small expected frequencies. Items differ in difficulty, just as subjects differ in ability. These two facts will give rise to large deviations from the mean observation number of 5. Especially in the high and low scoring subsamples low expected frequencies are bound to occur.

It is known that small expected frequencies damage the stability of chi-square statistics. The following example stems from Van den Wollenberg (1979). In a $Q_2$ analysis of the ISI-tests he found the following contingency table of expected and observed frequencies:

| | observed | | | expected | | |
|---|---|---|---|---|---|---|
| | 0 | 1 | | 0 | 1 | |
| 1 | 5 | 305 | 310 | 6.86 | 303.14 | 310 |
| 0 | 2 | 8 | 10 | .14 | 9.86 | 10 |
| | 7 | 313 | 320 | 7 | 313 | 720 |

Although in an absolute sense the deviations between observed and expected frequencies were small, the small expected frequency .14 caused a contribution to $Q_2$ of 21.93. Van den Wollenberg (1982) discussed an instance with simulated data conforming the model, in which even the $Q_1$ statistic suffered from this instability. The number of items was only 8, the number of subjects was 4000, the number of cells was only 112, but the parameters were very extreme ranging from -4. to +4. Over a total of 100 replications a mean $Q_1$ of 43.31 was obtained, where the expectation was 42. This may still seem reasonable, but the observed variance was 283.35, whereas the theoretical value was 84. Indeed the source of these heavy deviations was lying

in the extreme score groups, some giving rise to very high values of the statistic. The mean value is less affected by extreme values than the variance, by which the results can be understood.

In principle all statistics mentioned in this paper may suffer from the indicated phenomenon. In fact in the same study the Martin Lof statistic proved to be more sensitive to small expected frequencies than the $Q_1$ statistic.

It may be obvious that the applicability of $Q_2$ would be greatly improved if the number of cells in the observation matrix could be drastically decreased.

Another problem in the application of the $Q_2$ statistic is the accuracy of computations. It is a well known fact that the difference algorithm for computing the elementary symmetric functions and their first order partial derivatives (e.g. Fischer, 1974) becomes inaccurate, when the number of items exceeds 20. In order to obtain the $Q_2$ statistic the second order partial derivatives are also needed adding another recursive stage and another source of computational inaccuracy.

Gustafsson (1980b) proposes an alternative algorithm for computing the elementary symmetric functions and their first order partial derivatives, which gives also accurate results for large values of k. In another contribution to the present issue Jansen (1981) shows how the second order probabilities can be obtained from the first order probabilities, which implies that the second order partial derivatives are not needed.

So it seems that the accuracy problem associated to Conditional Maximum Likelihood estimation and the statistics $Q_1$ and $Q_2$ have been solved.

## 4. $Q_2$ for composite partitionings

All the disadvantages of the $Q_2$ procedure mentioned in the preceding section would be reduced to a considerable degree, when it would prove possible to drop the requirement that the parameters should be estimated in every level group separately. In the present section we will study the possibilities to use other partitionings of the dataset than the one we used until now. In the first part the theoretical aspects are discussed, whereas in the second part simulation results are presented. In order to prevent confusion it seems useful to introduce some shorthands for the partitionings discussed in the present section. In the following we will use the terms:

| | |
|---|---|
| raw | the sample is divided into k-3 level groups, one group for each relevant raw score. |
| half | the sample is divided in a high- and a low-scoring subsample |
| three | the sample is divided in a high, a low and an intermediate score group |
| total | all level groups are taken into one 'subsample' |

### 4.1. Some theoretical considerations

The $Q_{2r}$ statistic is a sum of statistics $q_{rij}$ which are obtained within each level group for each item pair. For each item pair a 2*2 contingency table is obtained. As Van den Wollenberg (1979, p 127) points out, the observed marginals of this contingency table should be equal to the expected marginals. This requirement holds, when in each level group the item parameters are estimated. Molenaar (1980) studies this point very explicitely considering several cases. Only in the case, where level group parameters were used and the second order

probabilities were obtained conditional upon the marginals $n_{ri}$ and $n_{rj}$, the $q_{rij}$ had an asymptotical chi-square distribution. It is exactly this procedure that was used by Van den Wollenberg (1979,1982).

When the sample is divided into level groups and $q_{rij}$ statistics are obtained within level groups, it is necessary to obtain item parameter estimates for each level group separately. The study of Molenaar showed that overall estimates in combination with statistics within level groups do not give chi-square distributed statistics. However, there is yet another way to obtain the statistics.

When a sample is divided into level groups, these groups can be taken together into composite partitionings, as for instance a partitioning high-intermediate-low. Now item parameters are obtained within each of these subsamples and the $Q_2$ statistic is obtained for the whole subsample simultaneously. When we use g as a subsample index, the first and second order expected frequencies are obtained as

$$E(n_{gi}) = \sum_r n_r \frac{\hat{\varepsilon}_i^{(g)} \hat{\gamma}_{r-1}^{(i)}(\underline{\hat{\varepsilon}}^{(g)})}{\hat{\gamma}_r(\underline{\hat{\varepsilon}}^{(g)})} \qquad (r = g_1,\ldots,g_2) \quad . \tag{11}$$

$$E(n_{gij}) = \sum_r n_r \frac{\hat{\varepsilon}_i^{(g)} \hat{\varepsilon}_j^{(g)} \hat{\gamma}_{r-2}^{(i,j)}(\underline{\hat{\varepsilon}}^{(g)})}{\hat{\gamma}_r(\underline{\hat{\varepsilon}}^{(g)})}$$

Here $g_1$ is the lowest score in the subsample and $g_2$ is the highest score.

Now the following contingency tables can be obtained:

observed

| i \ j | 0 | 1 | |
|---|---|---|---|
| 1 | | | $n_{gi}$ |
| 0 | | | $n_{g\bar{i}}$ |
| | $n_{g\bar{j}}$ | $n_{gj}$ | $n_g$ |

theoretical

| i \ j | 0 | 1 | |
|---|---|---|---|
| 1 | | | $E(n_{gi})$ |
| 0 | | | $E(n_{g\bar{i}})$ |
| | $E(n_{g\bar{j}})$ | $E(n_{gj})$ | $n_g$ |

It can be easily seen that the expected and observed marginals are equal, because the item parameters are estimated by solving the the following set of equalities:

$$(12) \qquad n_{gi} = \sum_r n_r \frac{\hat{\varepsilon}_i^{(g)} \hat{\gamma}_{r-1}^{(i)}(\underline{\hat{\varepsilon}}^{(g)})}{\gamma_r(\underline{\hat{\varepsilon}}^{(g)})} \quad ,$$

which is equivalent to the equation used to obtain the first order expected frequencies in (11), when estimated parameters are used. The 2*2 contingency table follows an extended hypergeometric distribution (Molenaar (1980)). It can be shown that the requirements of Harkness( (1965), as cited in Molenaar (1980) are fulfilled, which implies that the $q_{rij}$ statistic is equal to the normal approximation of the extended hypergeometric. As a consequence each $q_{rij}$ is asymptotically distributed as chi-square with one degree of freedom (°).

(°) As the study of Molenaar is not yet officially published, we will at this moment abstain from an elaborate exposition on this point.

## 4.2. Distributional properties

When the individual $q_{rij}$ statistics are summed to obtain the statistic $Q_{2r}$, it is assumed that the covariances of the individual statistics are properly dealt with by means of the correction factor $(k-3)/(k-1)$. For the raw score partitioning this has been shown to be the case to a satisfactory degree.

In order to check whether the approximation still holds good for composite partitionings, a Monte Carlo study was performed. The data-construction procedure was the same as used by Van den Wollenberg (1982). A short outline is given here:

1 A number of k fixed item parameters are fed into the generation program.

2 A subject is sampled from the standard normal distribution

3 By means of the parameters under 1) and 2) the response probabilities are obtained according to the basic equation of the Rasch model.

4 A vector of k independent elements, one for each item, is sampled from the uniform distribution with domain (0,1).

5 The k probabilities under 3) are compared with the corresponding random numbers under 4). An item is said to be positively responded by the subject, when the probability exceeds the corresponding random number.

6 Steps 2-5 are repeated N times for desired sample size N.

For the present simulations we sampled 4000 subjects. In table 1 the results are presented for a series of simulations with equal item parameters. The number of items varied from 6 through 10. The mean values are based upon 100 replications each.

Table 1 The $Q_2$ statistic for the partitionings half, three and total; datasets conforming the Rasch model, equal item-parameters, N=4000, 100 replications.

| | number of items | | | | |
|---|---|---|---|---|---|
| | 6 | 7 | 8 | 9 | 10 |
| | | | TOTAL | | |
| df | 9 | 14 | 20 | 27 | 35 |
| $Q_2$ | 8.67 | 14.09 | 19.49 | 27.04 | 33.68 |
| $2*df$ | 18 | 28 | 40 | 54 | 70 |
| $s^2(Q_2)$ | 10.38 | 24.58 | 24.66 | 35.91 | 50.68 |
| $K(Q_2)$ | .101 | .072 | .107 | .055 | .119 |
| | | | HALF | | |
| df | 18 | 28 | 40 | 54 | 70 |
| $Q_2$ | 18.35 | 28.93 | 399.32 | 54.77 | 70.18 |
| $2*df$ | 36 | 54 | 80 | 108 | 140 |
| $s^2(Q_2)$ | 29.46 | 54.02 | 77.62 | 89.97 | 111.50 |
| $K(Q_2)$ | .077 | .106 | .095 | .080 | .091 |
| | | | THREE | | |
| df | 27 | 42 | 60 | 81 | 105 |
| $Q_2$ | 26.87 | 42.06 | 60.26 | 82.90 | 104.86 |
| $2*df$ | 54 | 84 | 120 | 162 | 210 |
| $s^2(Q_2)$ | 59.69 | 60.69 | 119.98 | 189.92 | 163.36 |
| $K(Q_2)$ | .060 | .065 | .034 | .108 | .074 |

some critical values for the Kolmogorov goodness of fit test statistic K:

| $\alpha$ | $K_\alpha$ |
|---|---|
| .20 | .107 |
| .10 | .122 |
| .05 | .136 |

From table 1 it becomes clear that the partitionings used here operate satisfactorily for equal item parameters. The mean values of the statistics are in all cases very near the expectations. The variance, of course, shows larger variability, tending to be a bit smaller than expected. The Kolmogorov statistic, K, for the deviation of observed distribution and theoretical chi-square distribution is very satisfactory indeed. In no instance the 5% significance level was reached. So it can be stated that for equal item parameters the the distribution of $Q_2$ is closely approximated by a $\chi^2$-distribution

In table 2 the same model tests are presented, now for item pools with differing item parameters ranging from -2. through +2. The results are definitely worse than for the case of equal item parameters, and in some instances clearly significant values of the Kolmogorov statistic were obtained. Still the approximation can be looked upon as a good one for the following reasons:

- Only a minority of fit tests shows significant values

- The sample sizes are very large; each systematic deviation from the model, however small, will become manifest.

- Such deviations from the theoretical chi-square distribution as there exist tend to make the test a bit on the conservative side. This seems a rather harmless property for a model test.

Table 2 The $Q_2$ statistic for high-low, three and total sample partitionings of datasets conforming the Rasch model; N=4000, unequal item parameters ranging from -2. through +2.

| | number of items | | | | |
|---|---|---|---|---|---|
| | 6 | 7 | 8 | 9 | 10 |
| | | | TOTAL | | |
| $\underline{df}$ | 9 | 14 | 20 | 27 | 35 |
| $Q_2$ | 8.75 | 13.83 | 17.66 | 26.66 | 32.60 |
| $2*\underline{df}$ | 18 | 28 | 40 | 54 | 70 |
| $s^2(\underline{Q_2})$ | 11.80 | 17.76 | 22.19 | 31.94 | 40.45 |
| $K(Q_2)$ | .086 | .068 | .181 | .137 | .188 |
| | | | HALF | | |
| $\underline{df}$ | 18 | 28 | 40 | 54 | 70 |
| $Q_2$ | 16.81 | 27.98 | 38.69 | 54.44 | 67.66 |
| $2*\underline{df}$ | 27 | 42 | 60 | 81 | 105 |
| $s^2(\underline{Q_2})$ | 32.18 | 47.65 | 71.85 | 80.76 | 88.50 |
| $K(Q_2)$ | .131 | .087 | .137 | .064 | .137 |
| | | | THREE | | |
| $\underline{df}$ | 27 | 42 | 60 | 81 | 105 |
| $Q_2$ | 25.82 | 41.81 | 58.57 | 82.86 | 103.79 |
| $2*\underline{df}$ | 54 | 82 | 120 | 162 | 210 |
| $s^2(\underline{Q_2})$ | 54.66 | 97.87 | 126.50 | 167.18 | 149.93 |
| $K(Q_2)$ | .114 | .063 | .144 | .074 | .093 |

The present results bear ressemblence to the results obtained for the raw score partitioning (Van den Wollenberg, 1982). For equal item parameters the fit is extraordinary, for unequal parameters it is less but still satisfactory.

The conclusion from the present simulations is that other partitionings of the dataset than the raw score partitioning into k-3 level groups also lead to statistics which can be approximated by $\chi^2$ to a sufficient degree. It is even admissible to take the whole sample together and obtain only one set of item parameter estimates and compute $Q_2$ for the whole relevant sample.

## 4.3. The detection of multidimensionality

In the preceding section we have seen that the chi-square distribution approximates the distribution of $Q_2$ to a satisfactory degree, even when the sample is partitioned into less than k-3 level groups. Now we will inspect whether the statistic in conjunction with alternative partitionings is also sensitive to violation of the dimensionality axiom and to what extent. In table 3 the results are presented of single runs on datasets violating the dimensionality axiom. This was done by sampling two parameters for each subject from the bivariate standard normal distribution with correlation zero between variates and letting some items appeal to one subject parameter and other items to the other (see also Van den Wollenberg (1979, 1982)). It may be observed that the alternative partitionings are also sensitive to violation of the model.

Table 3 $Q_2$ statistics for several partitionings of the dataset; two-dimensional latent space, N=4000.

a) equal item parameters

| | number of items | | | | |
|---|---|---|---|---|---|
| | 6 | 7 | 8 | 9 | 10 |
| total | 332.49 | 516.48 | 611.23 | 774.70 | 1080.68 |
| | df= 9 | df= 14 | df= 20 | df= 27 | df= 35 |
| half | 401.62 | 608.88 | 775.03 | 963.85 | 1355.87 |
| | df= 18 | df= 28 | df= 40 | df= 54 | df= 70 |
| three | 441.71 | 552.68 | 841.64 | 1021.91 | 1469.57 |
| | df= 27 | df= 42 | df= 60 | df= 81 | df=105 |
| raw | 441.71 | 576.81 | 929.62 | 1266.63 | 1715.90 |
| | df= 27 | df= 56 | df=100 | df=162 | df=245 |

b) item parameters ranging from -2. through 2.

| | | | | | |
|---|---|---|---|---|---|
| total | 208.79 | 300.43 | 445.28 | 587.03 | 616.41 |
| | df= 9 | df= 14 | df= 20 | df= 27 | df= 35 |
| half | 290.87 | 345.21 | 584.70 | 693.52 | 798.55 |
| | df= 18 | df= 28 | df= 40 | df= 54 | df=70 |
| three | 344.69 | 422.87 | 648.64 | 742.84 | 895.38 |
| | df= 27 | df= 42 | df= 60 | df= 81 | df=105 |
| raw | 344.69 | 451.76 | 712.59 | 876.88 | 1076.60 |
| | df= 27 | df= 56 | df=100 | df=162 | df=245 |

The power of the test is higher, when all item parameters are equal, which can be accounted for by the fact that in this case (item parameters equal to the mean of the subject distribution) the amount of statistical information in the dataset is larger. The differences between the several tests are not very large, especially not when it is recognized that the more elaborate partitionings have more degrees of freedom. So when extreme expected frequencies prohibit the use of the complete raw score partitioning, which will almost always be the case, other partitionings may be used instead.

The partitioning of the dataset into three level groups high-intermediate-low turns out to be the best one of the present composite

partitionings. One reason for this is obvious: this partitioning is the most elaborate one; the deviations are accumulated over more instances. There is yet another reason.

As Van den Wollenberg (1979) argues, the high and low scoring groups will not be very helpful in the detection of violation of unidimensionality. High scoring subjects tend to have high parameters on both latent traits and therefore a high scoring subsample is homogeneous with regard to the underlying traits. The same argument holds for low scoring subsamples. The intermediate score groups will contain subjects scoring high on one trait and low on the other (or the other way around) and subjects that score intermediate on both traits. In other words the intermediate groups are heterogeneous with respect to the latent traits and this will show in lack of local stochastic independence.

The above is illustrated by listing the chi-square contributions of the level groups to the total statistic. Below this is done for the total partitioning of the 10 item, equal parameter case:

| level group | $Q_{2r}$ (df=35) |
|---|---|
| 2 | 39.22 |
| 3 | 117.25 |
| 4 | 185.37 |
| 5 | 345.29 |
| 6 | 253.42 |
| 7 | 101.99 |
| 8 | 34.05 |

It is seen that violation manifests itself especially in the intermediate groups. When these groups are taken together in the partitioning, deviations will accumulate and violation becomes manifest more easily.

From table 2 it also seems that for the partitioning 'three' the approximation of the chi-square distribution is better than, for

instance, for the total partitioning. The partitioning 'three' seems to be an attractive one to use standardly in connection with the $Q_2$ statistic.

## 5. Conclusions

The Rasch model is an important advancement in the measurement of individual differences. The model makes it possible to compare persons in an objective way, once it has been concluded that the model holds for a given universe of subjects and a universe of measurement devices, say items.

Except for a special model test introduced by Martin Lof (1973), which is another test than the better known test statistic T of Martin Lof, all model testing has until recently been concentrated on the equality of item parameters over subsamples. Equality of item parameters over subsamples is a necessary condition for the model to hold, but it is not a sufficient condition, as was explicitely demonstrated by Van den Wollenberg (1979,1982). The violation of local stochastic independence and unidimensionality may be overlooked by the traditional test procedures.

The $Q_2$ statistic is especially sensitive to violation of these two axioms, and thus fills the existing gap. However, the application of the statistic was not without problems, as item parameters had to be estimated in every level group, which entailed several serious problems.

In the present study it was demonstrated that other partitionings than the complete raw score one are possible. Use of one of the discussed alternatives implies:

- The number of subsamples decreases and with it the number of times the item parameters have to be estimated, implying a gain in computing time.

- The sizes of the subsamples increase, which garantees that, under normal circumstances, the item parameters can always be estimated.

- The number of cells in the observation matrix reduces considerably, so the mean number of observations becomes larger, which increases the stability of the statistics $q_{rij}$.

The partitioning can be choosen depending upon the circumstances. For large datasets and for initial runs the 'total' partitioning may be used, implying only 2k(k-1) cells in the observation matrix and $N^{*}\frac{1}{2}k(k-1)$ observations. For this partitioning only one set of item parameters has to be estimated and all relevant subjects are included in the same 'subsample'. For more intensive testing of the model, the more elaborated partitionings can be used to the extent allowed by the idiosyncrasy of the dataset at hand.

Given the above arguments, it does not seem too optimistic to state that by the present results, $Q_2$ has become a statistic, which can be applied to most datasets without problem.

## REFERENCES

Andersen, E.B. A goodness of fit test for the Rasch model. Psychometrika, 1973, 38, 123-140.

Fischer, G.H. Einfuehrung in die Theorie psychologischer Tests. Bern: Huber, 1974.

Fischer, G.H. & Scheiblechner, H.H. Algorithmen und Programmen fuer das probabilistische Testmodel von Rasch. Psychologische Beitraege, 1970, 12. 23-51.

Gustafsson, J.E. Testing and obtaining fit of data to the Rasch model. British Journal of Mathematical and Statistical Psychology, 1980, 32 205-233. (a)

Gustafsson, J.E. A solution of the conditional estimation problem for long tests in the Rasch model for dichotomous items. Educational and Psychological Measurement, 1980, 40, 377-385. (b)

Gustafsson, J.E. & Lindblad, T. The Rasch model for dichotomous items: A solution estimation problem for long tests and some thoughts screening procedures. Paper presented at the European Conference on Psychometrics and Mathematical Psychology, Uppsala, June 15-17, 1978.

Harkness, W.L. Properties of the extended hypergeometric distribution. Annals of Mathematical statistics. 1965, 938-945.

Jansen, P.W.G. Herschrijving van de verwachte proportie $p_{ij}$ in het Rasch model (submitted to KM)

Martin Lof, P. Statisika modeller. Anteckningar fran seminarier lasaret 1969-1970 utarbetade av Rolf Sunberg obetydligt andrat nytryk, oktober Stockholm: Institutet for Forsakringsmatematik och Matematisk Statistik vid Stockholms Universitet, 1973.

Molenaar, I.W. Some improved diagnostics for failure of the Rasch model. Heymans Bulletin 80-482-EX, 1980.

Stelzl, I. Ist der Modelltest des Rasch-Modells geeignet, Homogenitaets Hypothese zu pruefen? Ein bericht ueber Simulation Studien mit inhomogene Daten. Zeitschrift fuer experimentelle und angewandte Psychologie, 1979, XXVI, 652-672.

Van den Wollenberg, A.L. The Rasch model and time limit tests. dissertation, Nijmegen, 1979.

Van den Wollenberg, A.L. On the Wright-Panchapakesan goodness of fit test for the Rasch model. Internal Report MA-80-02, University of Nijmegen, 1980.

Van den Wollenberg, A.L. A simple and effective method to test the dimensionality axiom of the Rasch model. Applied Psychological Measurement, 1981, in press.

Van den Wollenberg, A.L. Two new test statistics for the Rasch model. Psychometrika, 1982, in press.

Wright, B.D. & Panchapakesan, N. A procedure for sample-free item analysis. Educational and Psychological Measurement, 1969, 28, 229-248.