PSYCHOMETRIKA—VOL. 48, NO. 1
MARCH, 1983

SOME IMPROVED DIAGNOSTICS FOR FAILURE OF THE RASCH MODEL

IVO W. MOLENAAR

UNIVERSITY OF GRONINGEN

Although several goodness of fit tests have been developed for the Rasch model for dichotomous items, most of them are of a global, asymptotic, and confirmatory type. This paper, based on ideas from a recent thesis by Van den Wollenberg, offers some suggestions for local, small sample, and exploratory techniques: difficulty plots for person groups scoring right and wrong on a specific item, a slope test per item based on a binomial distribution per score group, and a unidimensionality check based on an extended hypergeometric distribution per score group.

Key words: logistic latent trait model, Rasch assumptions, exploration of model fit.

1. Introduction

Let n subjects, each characterized by a latent subject parameter θ_v ($v = 1, 2, \ldots, n$), give answers to k dichotomous items, each characterized by an easiness parameter ε_i ($i = 1, 2, \ldots, k$). Under the four postulates of monotonicity, sufficiency of total number of items correct for the subject parameter, unidimensionality, and local independence, the probability of a positive response is given by

$$P(A_{vi} = 1) = \frac{\theta_v \varepsilon_i}{1 + \theta_v \varepsilon_i}. \tag{1.1}$$

For a full discussion of the logical equivalence of (1.1) and the four postulates see Fischer [1974, Ch. 12]; it is important that up to trivial reparametrizations only (1.1) satisfies the postulates. Properties of this so-called Rasch model, or one-parameter logistic model, have been investigated for the past twenty years: for reviews see e.g., Fischer [1974], Wright and Stone [1979], Van den Wollenberg [1979], Wainer, Morgan and Gustafsson [1980]. Several computer programs for conditional maximum likelihood estimation of the item parameters are available, cf. Fischer [1974], Allerup and Sorber [Note 1], Gustafsson [Notes 2 & 3], Raaijmakers and Van den Wollenberg [Note 4].

The postulates mentioned above are a stumbling block against regular application of the Rasch model in two respects. First, they deal with aspects of latent quantities, and thus are difficult to check by statistical tests based on observed data for a random sample of persons. Second, they are rather strong assumptions, in many cases not valid for the total data matrix but only for a subset of items and or persons as yet to be selected. There is a need for goodness of fit tests for the model which can simultaneously serve as diagnostic aids in such a selection process.

Tests have been proposed among others by Andersen [1973; 1982], Fischer and Scheiblechner [1970], Martin-Löf [Note 5], Wright and Panchapakesan [1969], and Stene [Note 6]. In their introduction to the EBA computer program, Allerup and Sorber

This paper owes much to the inspiring and pioneering work of Arnold Van den Wollenberg, of which only minor aspects are criticized. Thanks go to Charles Lewis for stimulating discussions and for solutions to some programming problems.

Requests for reprints should be sent to Ivo W. Molenaar, Fac. Soc. Wet., Oude Boteringestraat 23, 9712 GC Groningen, Netherlands.

0033-3123/83/0300-3070$00.75/0

© 1983 The Psychometric Society

[Note 1] list a few more. Recently Van den Wollenberg [1982] added two more tests, which will be discussed here.

Gustafsson [1980] presents a review of the problems involved in using the tests, of which several are incorporated in his PML program [Gustafsson; Notes 2 & 3].

Our discussion will center around the following themes:

(i) The use of exploratory rather than confirmatory techniques will suit the needs of many users: although guards against chance capitalization should be provided, formal significance tests are of limited value only.

(ii) A test should give specific information regarding the nature of a bad fit (which postulate, which item(s) or which persons contribute most?).

(iii) An exploratory tool for detecting Rasch homogeneous subsets of items is very useful.

(iv) Unless the data are already close to fulfilling the Rasch postulates the use of the total score per person for subdivision into homogeneous groups is inadequate.

(v) Once a reasonably good Rasch scale has been found, local tests per score group per item (or per item pair) should preferably use the exact rather than the asymptotic distribution of the test statistic.

2. *Exploration by Splitting on an Item*

Most analyses of the Rasch postulates are based on the principle of specific objectivity, which implies that essentially the same item difficulty estimates should be obtained for any subdivision of the persons into two or more groups. If an external criterion is available (e.g., sex or age of the person taking the items) for which a violation of the principle is a realistic alternative, the well known Andersen likelihood ratio test should be carried out after subdivision based on the criterion. It goes without saying that a main effect of the criterion, reflecting unequal ability distribution in the subgroups, is well incorporated in the model; a criterion × item interaction, on the other hand, violates the principle. A thorough discussion of this point is, e.g., found in Fischer [1974].

Subdivision on the basis of an internal criterion (i.e., obtained from the results on the k items to be analyzed) is a rather common alternative. Nearly always this is done on the basis of the total correct score, either in the form of a dichotomy of high and low scores, or in the form of an analysis per score group (concatenating low frequencies in most applications).

Referring back to our theme iv, we claim that this is a useful strategy when nearly all items do form a Rasch scale for one latent trait, but not when the item set considered is still rather heterogeneous. Van den Wollenberg [1979, p. 100] presents a theorem stating conditions under which the concatenation of two internally Rasch homogeneous subscales for different latent traits appears to be one homogeneous Rasch scale when the Andersen test is based on the total score.

Based on section 3.4.1 of Van den Wollenberg [1979] we propose to single out one item, here to be called the splitter, and to consider the two subgroups for which the answer to the splitter is zero and one, respectively. The difficulties of the remaining items are then estimated in both groups and plotted against each other (see also Formann, 1981).

While a more complete discussion will be presented after the presentation of some examples, we predict that items measuring the same latent trait as the splitter will be more difficult for the persons scoring zero on the splitter and more easy for persons scoring one. Items unrelated to this trait should in principle be equally easy for both groups; thus the graphical plot may enable us to distinguish between the two kinds of items.

For some examples we use a dataset in which 685 secondary school students replied to 20 items on their attitude towards school (items 1–10) and towards teachers (items 11–20); details on the data are available on request. We want to see whether our graphical plot can single out items having nothing to do with the latent trait measured by the other items. For this purpose we shall add to our 20 real items two fictitious ones: item 21 equals 1 if the respondent's sequence number is even and 0 otherwise; item 22 equals 1 if the sequence number is 343 or more and 0 otherwise.

The illustrations below plot the conditional maximum likelihood estimates, obtained via Gustafsson's PML program, of the item difficulties $\sigma_i = -\log \varepsilon_i$ for the two groups of persons in which the splitter is zero (horizontal axis) and one (vertical axis). In Figures 1–3 this is done for the 22 items just described, omitting the splitter in each case. This omission is necessary to prevent an artifact, see Van den Wollenberg [1979; p. 111–114]. The Andersen test for the two groups is also given, but we shall see that the plots contain more information, namely about the behavior of the individual items.

In Figure 1, where the meaningless item 21 is the splitter, the subgroups of persons with V21 = 0 and V21 = 1 respectively can be viewed as two random samples from the same population. Up to small sampling variations, the two estimates of item difficulty are thus equal, and indeed all item points lie very close to the main diagonal.

The situation is rather different in Figure 2 where the split is based on item 13. All items correlating positively with item 13 obtain on average lower scores, and thus higher difficulty estimates, in the V13 = 0 group than in the V13 = 1 group. Especially items 14 and 20, which have a high association with item 13, lie clearly below the main diagonal.

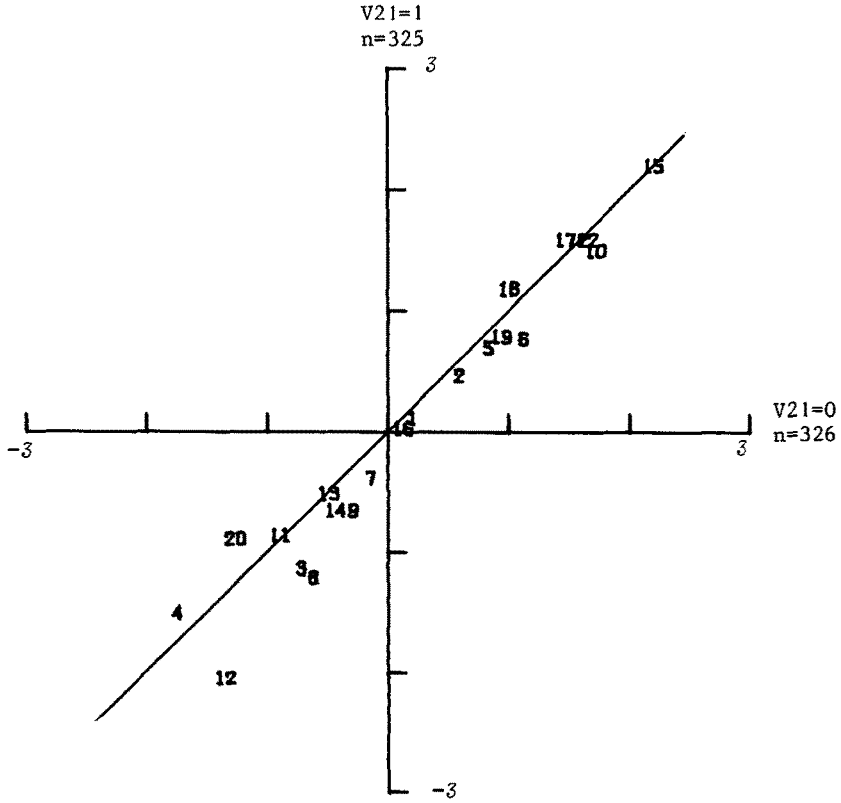

FIGURE 1
Plot of $\hat{\sigma}_i$ for items 1–22 with item 21 as splitter. Andersen test $\chi^2 = 19.5$, df = 20, $P = 0.49$.

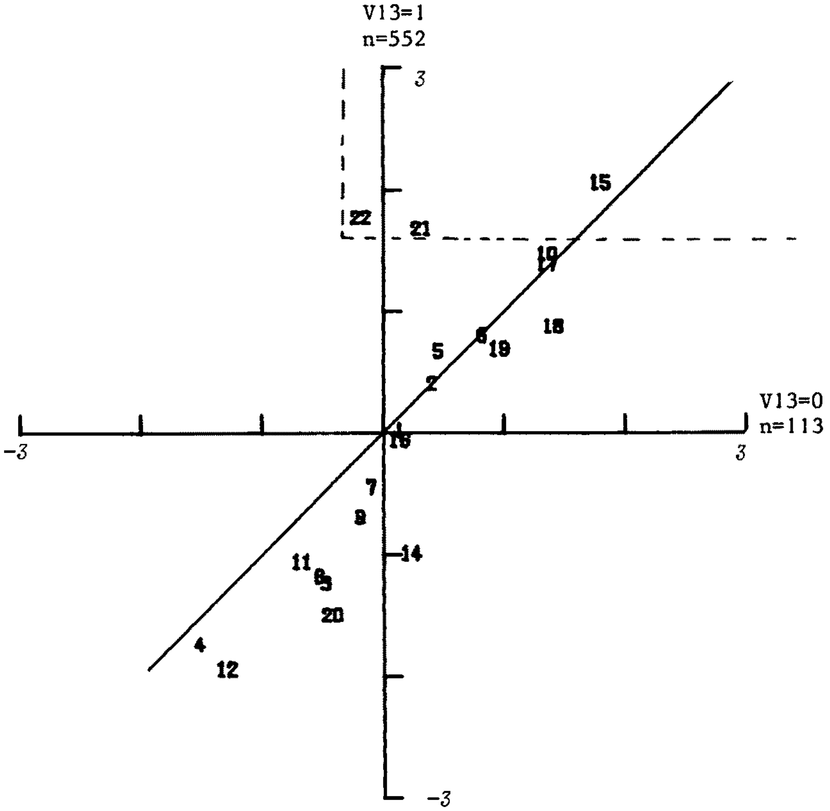

FIGURE 2
Plot of $\hat{\sigma}_i$ for items 1–22 with item 13 as splitter. The dotted lines indicate more meaningful coordinate axes, see text. Andersen test $\chi^2 = 165.3$, df $= 20$, $P < 0.00001$.

At first sight the position of items 21 and 22 is a little surprising. They seem to be of average difficulty in the V13 = 0 group, and rather difficult for the V13 = 1 group. As the positive attitudes towards school and teachers have nothing to do with the scores on items 21 and 22, we would have expected to find these items near the main diagonal, and all others clearly below and to the right of it.

This would be the case if the coordinate axes were given by the dotted lines in Figure 2. Such a translation of axes is permissible in the Rasch model: when the item parameters ε_i or $\sigma_i = -\log \varepsilon_i$ are estimated in one group of persons, one obtains a ratio scale for ε_i. It is well known that multiplication of all ε_i by the same factor, or addition of the same constant to all σ_i, leaves the model invariant as it can be compensated for in the subject parameters. It is customary to avoid this indeterminacy by a conventional norming condition for the item estimates, namely

$$\prod_{i=1}^{k} \hat{\varepsilon}_i = 1 \quad \text{or} \quad \sum_{i=1}^{k} \hat{\sigma}_i = 0. \tag{2.1}$$

As the PML program uses (2.1) for each subgroup separately, the sum of all x-coordinates in each plot is zero, and the same holds for the y-coordinates. Our outside knowledge that a difference in difficulty for items 21 and 22 is meaningless enables us to interpret the plot, using the dotted axes, as a sign that all other items are easier for the V13 = 1 group than for the V13 = 0 group.

Space does not permit us to exhibit all 22 plots of our item collection, but we may summarize that a split on item 22 leads to a plot resembling Figure 1, and a split on any

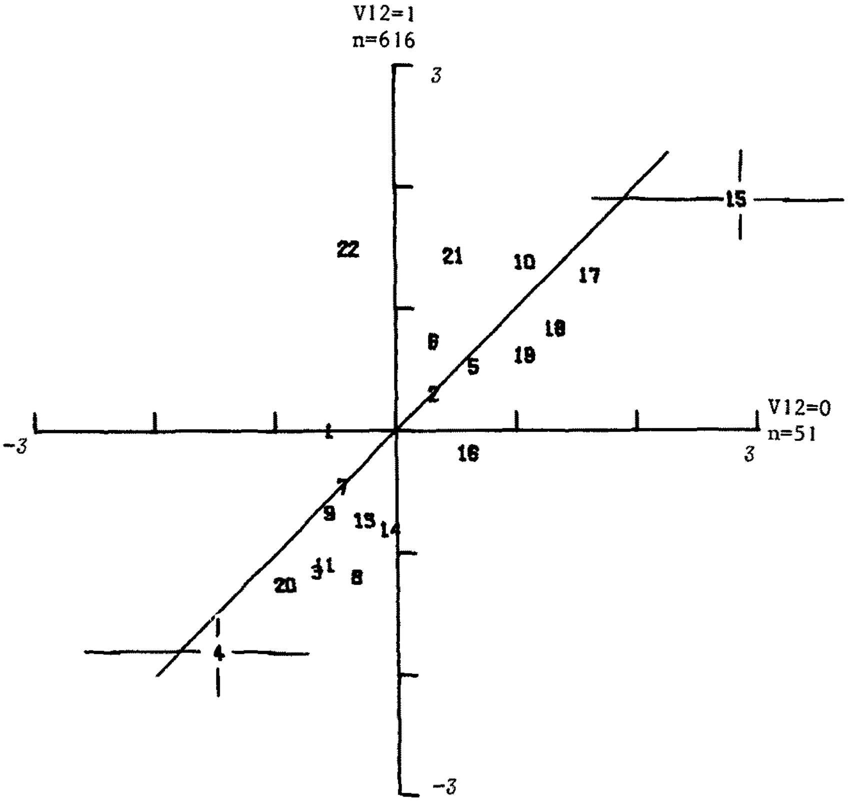

FIGURE 3
Plot of $\hat{\sigma}_i$ for items 1–22 with item 12 as splitter. Andersen test $\chi^2 = 69.6$, df = 20, $P < 0.00001$.

other item to a plot more or less resembling Figure 2. By "more or less" we refer to two reasons why a less satisfactory plot may appear, as illustrated in Figure 3.

With item 12 as the splitter, only 51 people belong to the V12 = 0 group, and 616 to the V12 = 1 group (the two numbers do not add up to 685, as persons with a zero or perfect total score cannot be used for estimation). In a group of 51 persons the item estimates are rather unreliable; as an indication the asymptotic 95 percent confidence intervals are plotted in Figure 3 for some items.

The second reason for a less satisfactory plot is that some of the 20 meaningful items may exhibit, by pure chance, a substantial correlation with items 21 and/or 22. As $r(12, 21) = 0.073$, item 12 is not the best splitter.

We conclude that it is wise to avoid the use of splitting items with very low or very high popularities, and to inspect plots for a few different splits, placing confidence only in the features that they have in common. In our example of 22 items, this would lead to the rejection of items 21 and 22, that consistently measure something not related to the trait underlying the other items. Our prior knowledge that these two items indeed measure no meaningful property is not necessary for our conclusion that they should be removed. Indeed they also will be rejected on the basis of the binomial plots discussed in the next section.

Some plots for the remaining 20 items are given in Figures 4 and 5. It appears that items 3, 9, and 6, and to a lesser extent 4, 8, 7, and 2, are positively related to the splitter item 10, and items 2, 3, 4, 7, 8, 9, 10 are positively related to the splitter item 6: they are more difficult for persons who responded zero to the splitter than to people responding 1.

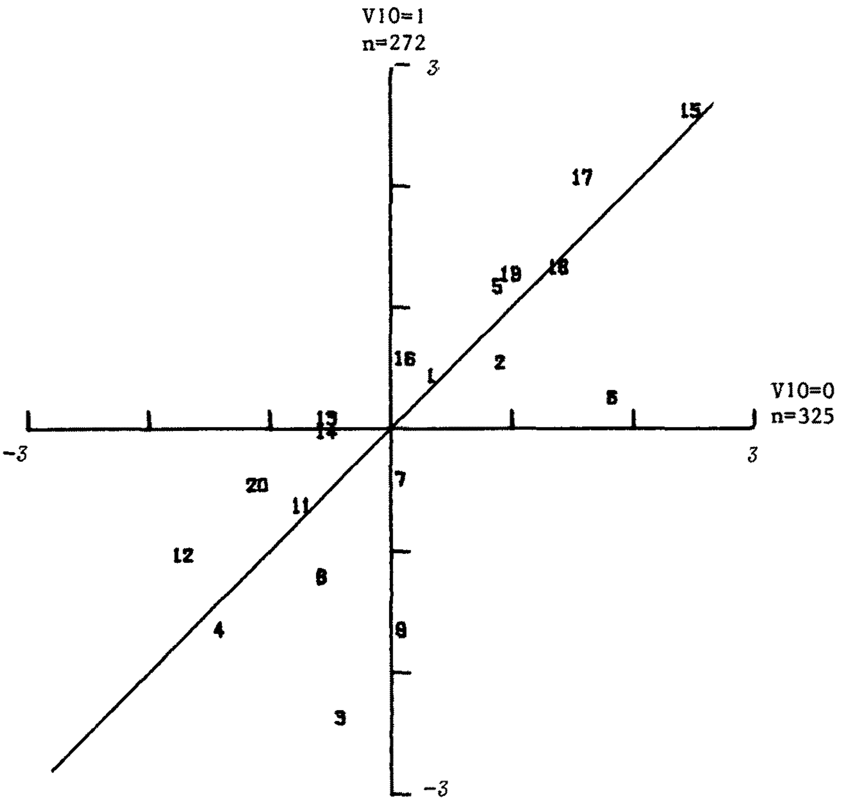

FIGURE 4
Plot of $\hat{\sigma}_i$ for items 1–20 with item 10 as splitter. Andersen test $\chi^2 = 146.9$, df = 18, $P < 0.00001$.

Although all items lie at a relatively close distance to the main diagonal, the plots suggest a closer relation between the items just mentioned. The items 11 to 20 (possibly also 1 and/or 5, and possibly not 18) seem to lie above and to the left of the main diagonal, which may again be an artifact of the convention (2.1).

Figures 6 and 7 illustrate that the reverse behavior occurs when the splitter is 17 or 19; note that 18 now seems to belong to the 11–20 scale, and 5 comes closer to the 1–10 scale. As the plots using other splits roughly yield the same conclusion and as the Andersen test based on total score indicates that the complete item set 1–20 is not a Rasch-scale, we conclude from the plots that meaningful subscales may very well center around 2, 3, 4, 7, 8, 9, 10, and around 11 to 20 respectively.

From the item contents it is clear that attitudes towards school and towards teachers may indeed be distinguished. Item 1 (I believe the lessons are worthwhile) may perform better in the teacher subscale (items 11–20). We shall continue our analysis of the subscales in the next section.

Summarizing, we conclude that plots based on an item split are attractive when a total scale does not conform to the Rasch assumptions, for the following reasons:

a) the total score, which is unreliable if it is partly based on items not belonging in the scale, is not used in the plots, whereas it is in most other testing and plotting procedures;
b) both isolated items unrelated to the others and internally homogeneous subscales can be detected.

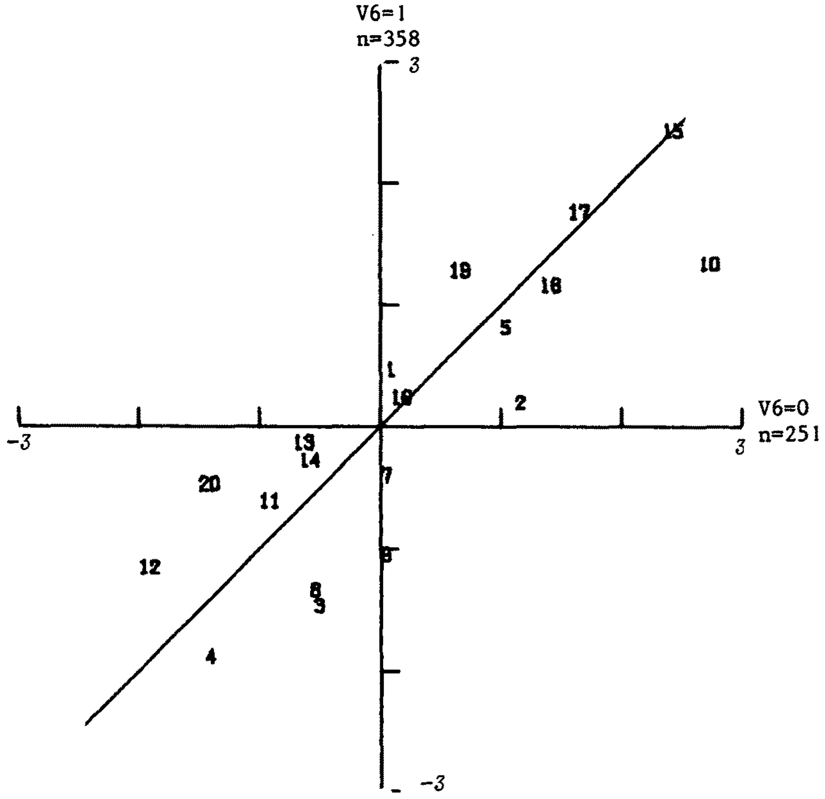

FIGURE 5
Plot of $\hat{\sigma}_i$ for items 1–20 with item 6 as splitter. Andersen test $\chi^2 = 140.4$, df = 18, $P < 0.00001$.

As a drawback, we mention that plots based on different splitters are not always similar, and that exploration by plots may lead to chance capitalization. Therefore, we shall discuss methods of a more inferential character in the next two sections.

3. Binomial Tests Per Item Score Group

3.1 Existing Tests

For this section, we shall need the following notation. Under the assumptions stated in the introduction, let n randomly chosen persons respond to k dichotomous items with success probabilities (1.1). Using capitals for random variables, let N_r be the number out of the total of n persons who have exactly r items correct ($r = 0, 1, \ldots, k$). Such persons are said to form the r-th score group. Within this group, let N_{ri} denote the number of persons who have item i correct ($i = 1, 2, \ldots, k$). It follows from our definition that trivially $N_{0i} = 0$ for all i and $N_{ki} = N_k$ for all i; score groups with zero or perfect score will not be considered any longer. If one conditions on $N_r = n_r$, as we shall do in the sequel, then N_{ri} has a binomial distribution with parameters n_r and

$$\pi_{ri} \overset{\text{def}}{=} \varepsilon_i \frac{\gamma_{r-1}^{(i)}}{\gamma_r}, \tag{3.1}$$

where the gamma symbols denote the elementary symmetric functions of the $\{\varepsilon_h\}$ in the usual way (see e.g. Fischer, 1974, p. 226 and 231).

Comparison of observed values for N_{ri} to their expected values $n_r \pi_{ri}$ (inserting esti-

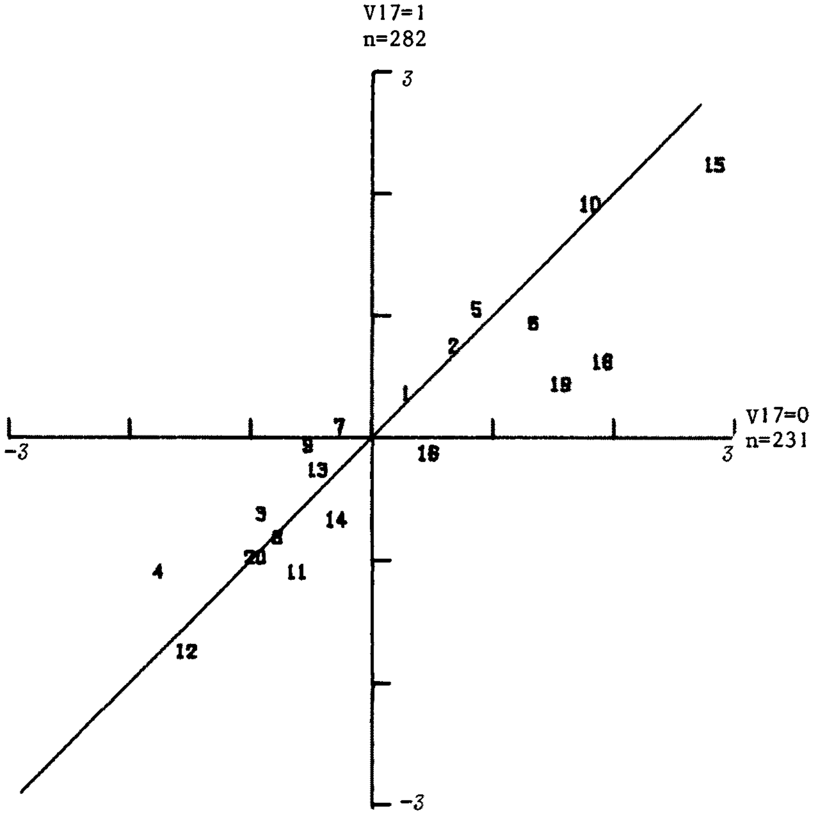

FIGURE 6

Plot of $\hat{\sigma}_i$ for items 1–20 with item 17 as splitter. Andersen test $\chi^2 = 101.4$, df = 18, $P < 0.00001$.

mated item difficulties) is attractive for several reasons. First, it avoids the iterative maximum likelihood estimation of item parameters in each subgroup that is necessary for the Andersen likelihood ratio test: in (3.1) one only inserts the estimates for the total group. Secondly, the individual comparisons show for what item and what score group the observed behavior possibly deviates from the Rasch model.

Binomial tests and plots are included in the computer programs EBA [Allerup & Sorber, Note 1] and PML [Gustafsson, Notes 2 & 3]. The combined statistic

$$T = \sum_{r=1}^{k-1} (\mathbf{N}_r - n_r \boldsymbol{\pi}_r)' V_r^{-1} (\mathbf{N}_r - n_r \boldsymbol{\pi}_r), \tag{3.2}$$

where

$$\begin{aligned} \mathbf{N}_r' &= (N_{r1}, N_{r2}, \ldots, N_{rk}) \\ \boldsymbol{\pi}_r' &= (\pi_{r1}, \pi_{r2}, \ldots, \pi_{rk}) \end{aligned} \tag{3.3}$$

and the $k \times k$ covariance matrix V_r has elements

$$\begin{aligned} &\frac{n_r \varepsilon_i \gamma_{r-1}^{(i)}}{\gamma_r} \quad \text{on the diagonal,} \\ &\frac{n_r \varepsilon_i \varepsilon_j \gamma_{r-2}^{(i,j)}}{\gamma_r} \quad \text{for } i \neq j, \end{aligned} \tag{3.4}$$

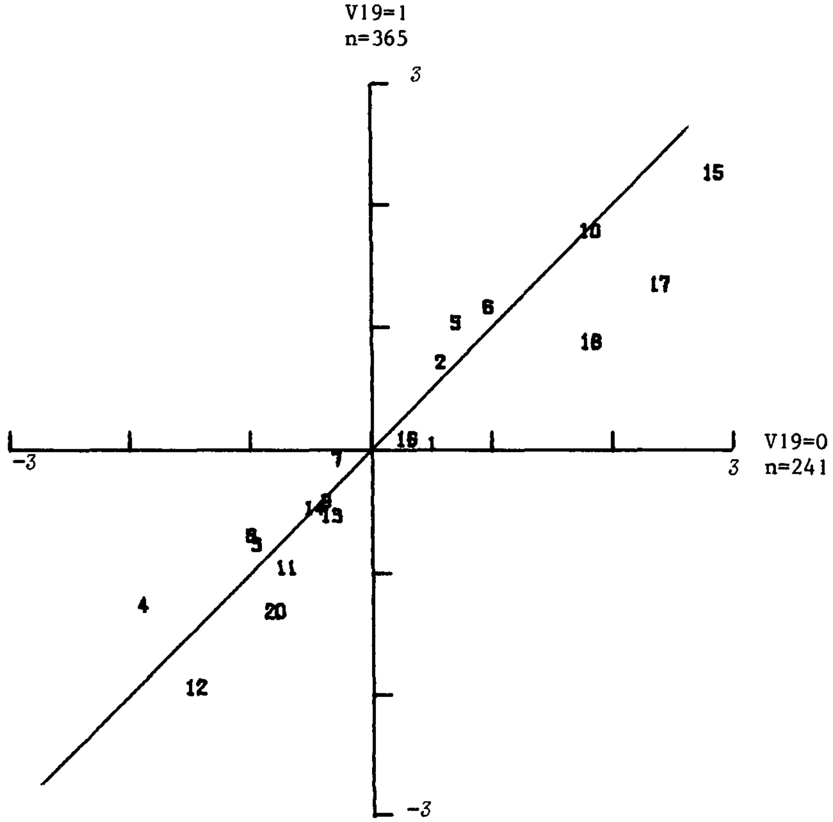

Figure 7
Plot of $\hat{\sigma}_i$ for item 1–20 with item 19 as splitter. Andersen test $\chi^2 = 73.4$ df = 18, $P < 0.00001$.

is used by Martin-Löf [Note 5, p. 128–129]. See also Gustafsson [1980], who includes this so-called ML–ICCSL test [T is asymptotically chi-square with $(k-1)(k-2)$ degrees of freedom when each $n_r \to \infty$] in his PML program, where he also prints terms of (3.2) for individual r.

Van den Wollenberg [1982, & 1979, p. 118–121] claims that

$$Z_{ri}^2 = \frac{(N_{ri} - n_r \pi_{ri})^2}{n_r \pi_{ri}} + \frac{[n_r - N_{ri} - n_r(1 - \pi_{ri})]^2}{n_r(1 - \pi_{ri})} \tag{3.5}$$

has a χ^2 distribution with one degree of freedom, if the true parameters are used for π_{ri} and n_r is assumed fixed. The claim is correct for $n_r \to \infty$, as (3.5) is equal to the square of

$$Z_{ri} = \frac{(N_{ri} - n_r \pi_{ri})}{\{n_r \pi_{ri}(1 - \pi_{ri})\}^{1/2}} \tag{3.6}$$

which as we saw is a standardized-binomial-random variable. Rather than using Martin-Löf's covariance matrix, he proposed the test statistic

$$Q_1 = \frac{k-1}{k} \sum_{i=1}^{k} \sum_{r=1}^{k-1} Z_{ri}^2, \tag{3.7}$$

which by the correction factor $(k-1)/k$ has approximately a χ^2 distribution with

$(k-1)(k-2)$ degrees of freedom if $\{\varepsilon_i\}$ are estimated from the data and the restriction

$$\sum_{i=1}^{k} N_{ri} = rn_r \tag{3.8}$$

is accounted for. In the special case when all items are of equal difficulty, or rather when in the sample

$$N_{.i} = \sum_{r=0}^{k} N_{ri} \tag{3.9}$$

has the same value for all i, the statistics T and Q_1 are identical; see Van den Wollenberg [1982] for a discussion of this case.

3.2 New Proposals

All authors quoted seem to concentrate their attention on the null hypothesis that the Rasch model is correct. The asymptotic chi-square test is powerful against any alternative set of probabilities π_{ri} not equal to (3.1). There are two special alternatives that we want to consider in detail.

When one item with index i is unrelated to the remaining ones, the probability of a correct answer to it in score group r exhibits only a minor increase with r for this index i, caused by the contribution of the item to the total score. More generally, we may expect

$$\begin{aligned} N_{ri} &> n_r \pi_{ri} \quad \text{for small } r \\ N_{ri} &< n_r \pi_{ri} \quad \text{for large } r \end{aligned} \tag{3.10}$$

for an item with a less steep trace line than the remaining items. A too steep trace line will result in reversed signs in (3.10). Inspection of the individual binomial plots or tests for each item in turn will thus reveal important information as to the nature of a violation of the Rasch assumptions when the inequalities (3.10) or their reverse hold for certain items. Such systematic deviations, which are frequently found, are more easily interpreted than an incidental deviation or an oscillation of N_{ri} around their expected value. Note that

$$\sum_{r=1}^{k-1} N_{ri} = N_{.i} - n_k = \sum_{r=1}^{k-1} n_r \pi_{ri} \tag{3.11}$$

if the item estimates $\hat{\varepsilon}_i$ based on the sufficient statistics $N_{.i}$ are used in (3.1). Thus any comparison between N_{ri} and $n_r \pi_{ri}$ results in the same sum when all score groups are used.

The BINO option in the PML computer program lists per item, for each score group for which n_r exceeds a user-specified minimum value (default = 5), the quantities n_r, N_{ri}, N_{ri}/n_r and π_{ri}, followed by the one-sided probability of exceedance

$$P_{ri} = \sum_{j=N_{ri}}^{n_r} \binom{n_r}{j} \pi_{ri}^j (1-\pi_{ri})^{n_r-j} \quad \text{when } N_{ri} > n_r \pi_{ri}, \tag{3.12}$$

$$P_{ri} = \sum_{j=0}^{N_{ri}} \binom{n_r}{j} \pi_{ri}^j (1-\pi_{ri})^{n_r-j} \quad \text{when } N_{ri} < n_r \pi_{ri}. \tag{3.13}$$

For an example we return to the 22 items discussed in section 2. Table 1 gives the results for item 21 when BINO was applied to all 22 items. The results for item 22 are similar, and no other item produced more than two P_{ri} values below 0.025 (corresponding to $\alpha = 0.05$ two-sided). It is clear that items 21 and 22 are correctly inferred to show too little relationship with the remaining items. For an easy comparison, the following signs

Table 1. Output of the BINO option in the PML program for an analysis of all 22 items; results for item 21.

score group	total freq.	number correct	observed		predicted	one-sided (3.12, 3.13)
r	n_r	n_{ri}	n_{ri}/n_r		π_{ri}	P_{ri}
3	6	2	.333	>>>>	.024	.0080
4	8	5	.625	>>>>	.035	.0000
6	12	4	.333	>>>>	.063	.0051
7	6	4	.667	>>>>	.080	.0005
8	10	2	.200	>	.101	.2677
9	23	10	.435	>>>>	.125	.0002
10	20	10	.500	>>>>	.154	.0003
11	22	8	.364	>>	.187	.0392
12	28	17	.607	>>>>	.225	.0000
13	31	16	.516	>>>>	.270	.0031
14	42	16	.381	>	.321	.2495
15	47	24	.511	>>>	.380	.0462
16	65	29	.446	=	.446	.5475
17	60	30	.500	<	.521	.4329
18	71	30	.423	<<<<	.603	.0016
19	77	37	.481	<<<<	.694	.0001
20	81	45	.556	<<<<	.791	.0000
21	51	35	.686	<<<<	.894	.0000

are printed in table 1 and 2 between N_{ri}/n_r and π_{ri}: for the case $N_{ri} > n_r\pi_{ri}$:

$$\begin{aligned} = &\text{ when } P_{ri} > 0.50; \\ > &\text{ when } 0.10 < P_{ri} \leq 0.50; \\ >> &\text{ when } 0.05 < P_{ri} \leq 0.10; \\ >>> &\text{ when } 0.01 < P_{ri} \leq 0.05; \\ >>>> &\text{ when } P_{ri} \leq 0.01, \end{aligned} \tag{3.14}$$

and the reversed signs in case (3.13).

Table 2 gives the results for items 3, 4, and 7 in the item set 2, 3, 4, 7, 8, 9, 10. Although only a few individual binomial tests reach the one-sided significance level of 0.025, the printed signs indicate that item 3 has a steeper trace line than expected, and item 4 a less steep one. The signs for item 7 exhibit a random behavior that is confirmed by high P_{ri} values. The complete results show that item 9 behaves like item 3. Figures 8 and 9 show two split plots for the item set used in Table 2. The split on 10 confirms our conclusions that items 3 and 9 have steeper trace lines than the remaining ones, but the split on 2 is less informative.

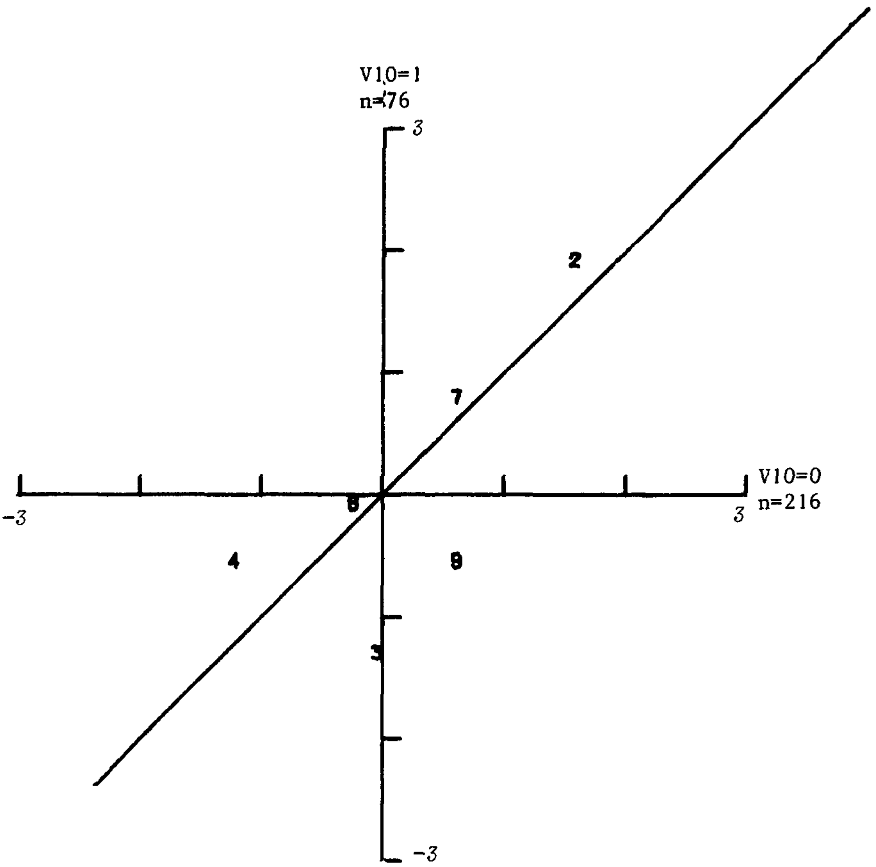

FIGURE 8
Plot of $\hat{\sigma}_i$ for items 2, 3, 4, 7, 8, 9, 10 with item 10 as splitter. Andersen test $\chi^2 = 21.8$, df = 5, $P = 0.0006$.

The idea that systematic deviations between N_{ri} and $n_r \pi_{ri}$ are the most plausible alternative to the binomial distributions predicted by the Rasch model, calls for a combination of one-sided binomial tests rather than a combination of all squared deviations into a statistic like T or Q_1. Moreover, it is informative to combine the evidence within one item across score groups rather than across items within a score group as advocated by Martin-Löf and Van den Wollenberg: an outlying item can be further studied and if necessary removed, but the information that a certain score group contributes much to the bad overall fit is hardly useful.

A combination of one-sided tests runs into the problem that some values of n_r may be large, making the test powerful, whereas other score groups contain so few individuals that the binomial test has little power and that any normal approximation may be inadequate. We suggest a combination procedure based on an *a priori* division of the index set of all score groups into two extreme classes for which different signs can be expected under meaningful alternatives, and a middle class where no relation between N_{ri} and $n_r \pi_{ri}$ is predicted. Let

$$n_m = n - n_0 - n_k, \tag{3.15}$$

and define r_1 and r_2 by

$$\sum_{r=1}^{r_1 - 1} n_r < \frac{n_m}{4} \leq \sum_{r=1}^{r_1} n_r, \qquad \sum_{r=r_2}^{k-1} n_r \geq \frac{n_m}{4} > \sum_{r=r_2+1}^{k-1} n_r. \tag{3.16}$$

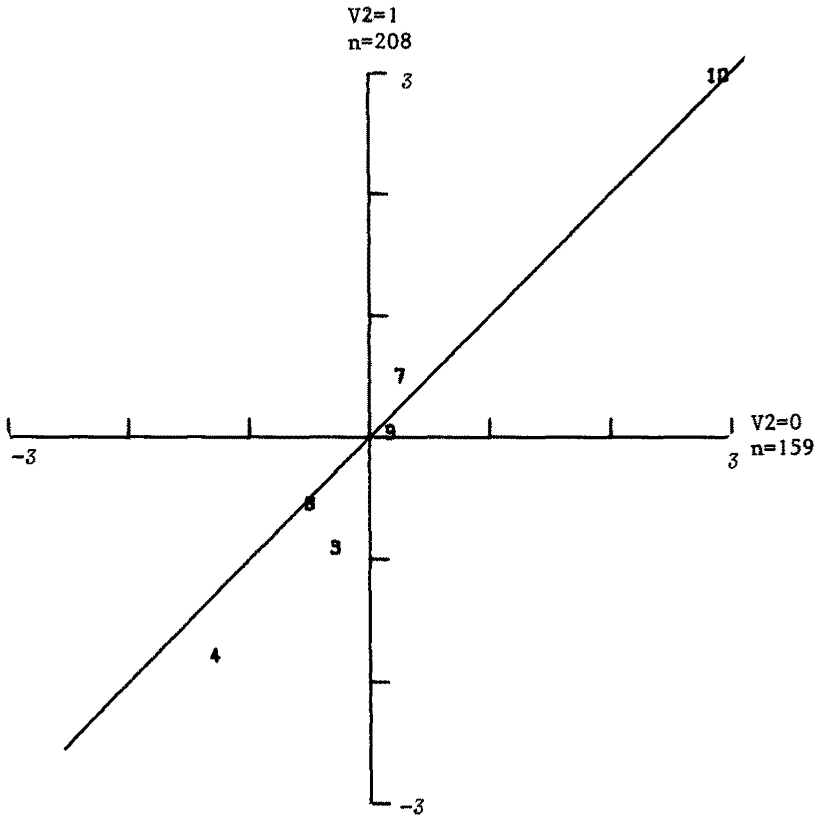

FIGURE 9
Plot of $\hat{\sigma}_i$ for items 2, 3, 4, 7, 8, 9, 10 with item 2 as splitter. Andersen test $\chi^2 = 5.1$, df $= 5$, $P = 0.41$.

It will be convenient to use the notation

$$\begin{aligned} L &= \{1, 2, \ldots, r_1\}, \\ M &= \{r_1 + 1, r_1 + 2, \ldots, r_2 - 1\}, \\ R &= \{r_2, r_2 + 1, \ldots, k - 1\}, \end{aligned} \tag{3.17}$$

for an exclusive and exhaustive partition of our index set of score groups (an empty set M is permitted, but an exceptional case in which (3.16) leads to $r_1 \geq r_2$ calls for revision).

Recall that Z_{ri} was introduced in (3.6) as the standardized version of the binomial variable N_{ri}. Our new proposal is the test statistic

$$U_i = \frac{\sum_L Z_{ri} - \sum_R Z_{ri}}{(r_1 + k - r_2)^{1/2}}. \tag{3.18}$$

Large positive values of U_i indicate a trace line that is not steep enough, or even an item unrelated or negatively related to the remaining ones. Such an item is a candidate for removal from the scale. Large negative values of U_i are an indication that item i measures the latent trait with a steeper trace line than the other items. Although the item may be "too good for the scale", its use in the unweighted total score may not be disastrous. Moreover, its position in the scale may well improve when items with very positive U_i are removed.

Table 2. Output of the BINO option in the PML program for an analysis of items 2, 3, 4, 7, 8, 9, 10; results for items 3, 4, and 7.

	score group	total freq.	number correct	observed		predicted	one-sided (3.12, 3.13)
	r	n_r	n_{ri}	n_{ri}/n_r		π_{ri}	P_{ri}
item 3	1	28	3	.107	<	.180	.2315
	2	30	7	.233	<<<	.405	.0384
	3	27	14	.519	<	.617	.1957
	4	37	31	.838	>	.786	.2952
	5	117	110	.940	>	.907	.1387
	6	165	165	1.000	>>>	.975	.0162
item 4	1	28	16	.571	>	.439	.1105
	2	30	21	.700	>	.676	.4756
	3	27	22	.815	=	.819	.5579
	4	37	33	.892	<	.907	.4555
	5	117	112	.957	<	.961	.4760
	6	165	160	.970	<<<	.990	.0271
item 7	1	28	3	.107	>	.074	.3440
	2	30	5	.167	<	.188	.4938
	3	27	12	.444	>	.346	.1898
	4	37	15	.405	<<	.551	.0533
	5	117	91	.778	<	.782	.4941
	6	165	158	.958	>	.940	.2194

We claim that U_i has under H_0 asymptotically (all $n_r \to \infty$) almost a standard normal distribution, the numerator being the sum of $r_1 + k - r_2$ terms with mean zero and variance one. The argument would be compelling if the Z_{ri} for different score groups were independent. By the restriction (3.11) and the insertion of estimated item parameters into π_{ri}, they are not. If we put

$$\begin{aligned} \mu_{ri} &= n_r \pi_{ri} \\ \tau_{ri}^2 &= n_r \pi_{ri}(1 - \pi_{ri}), \end{aligned} \tag{3.19}$$

the numerator of (3.18) is of the form

$$Y = \sum_{r=1}^{k-1} a_{ri}(N_{ri} - \mu_{ri}) \qquad \text{with} \qquad a_{ri} = \begin{cases} \tau_{ri}^{-1} & \text{for } r \in L \\ 0 & \text{for } r \in M \\ -\tau_{ri}^{-1} & \text{for } r \in R \end{cases} \tag{3.20}$$

The conditional variance of Y given that $V = \sum_{r=1}^{k-1} N_{ri}$ is by (3.11) assumed to be fixed, is known for the case when Y and V have a bivariate normal distribution (which is approximately true) to equal

$$\sigma_Y^2(1 - \rho_{YV}^2) = (r_1 + k - r_2) - \frac{\left\{\sum_L \tau_{ri} - \sum_R \tau_{ri}\right\}^2}{\sum_{r=1}^{k-1} \tau_{ri}^2}, \tag{3.21}$$

because cov $(Y, V) = \sum_{r=1}^{k-1} a_{ri}\tau_{ri}^2$ can be simplified for the weights a_{ri} given by (3.20). Because of the partial cancellation of positive and negative terms, the total expression (3.21) usually is close to $(r_1 + k - r_2)$. As the conditional variance is smaller, it is conservative to treat U_i as standard normal. In the author's experience, the difference is only nonnegligible when π_{ri} is very close to 1 for all $r \in R$ and not close to 0 for all $r \in L$, or vice versa, and/or when there is a rough bound in (3.16) due to the presence of a large score group. One could of course use (3.21) in the denominator of (3.18).

When U_i contains some terms for which n_r is small or moderate, say less than 30, then the corresponding Z_{ri} may be far from normally distributed under the null hypothesis, especially when simultaneously π_{ri} is close to zero or one. One might feel tempted to use other transformations of N_{ri} to normality than (3.6), or to replace Z_{ri} by the inverse normal cdf of the exact probability of exceedance P_{ri}. One could also use Fisher's omnibus procedure by which $\sum_r (-2 \log P_{ri})$ is asymptotically χ^2 under H_0, using for P_{ri} the binomial probability of at least N_{ri} for $1 \leq r \leq r_1$ and of at most N_{ri} for $r_2 \leq r \leq k - 1$, or for items with a too steep trace line just the reverse. It seems to us that none of these sophistications offers a satisfactory solution to the basic problem that both Z_{ri} itself and any transformation of it have a strictly discrete distribution if n_r is small (say $n_r < 10$, cf. Wallis, 1942).

We have also studied a test statistic based on the raw difference $\sum_L N_{ri} - \sum_R N_{ri}$. In many cases the most extreme score groups will simultaneously have low n_r values and large deviations between N_{ri}/n_r and π_{ri}. The new statistic would give such score groups less weight than groups with much larger n_r, whereas all groups involved in the summation get equal weight in U_i.

We are thus led to prefer U_i for an approximate significance test which is powerful against items for which the success probability as a function of r systematically deviates from π_{ri} in such a way that (3.10) or its reverse holds. To this end, Z_{ri} are printed for all r in our local modification of the BINO option in PML, and U_i is given for each item. If some values of n_r fall below the user-specified bound (default = 5), U_i can be modified to exclude the corresponding terms both in the summation and in the denominator. Other feasible modifications are the concatenation of adjoining score groups with small n_r, and the use of a significance level accounting for the number of significance tests involved in examining a complete item set. Since satisfactory experience has thus far been collected on the adequacy of the test based on U_i and on the choice of $n_m/4$ in (3.16), the calculation of U_i may become part of the standard analysis of an item set.

Let us summarize the results of this section. The comparison of individual observed N_{ri} values with the expected values $n_r \pi_{ri}$ under the Rasch model should only be used when the graphical analysis has given support to the suggestion that the total score of the item set under consideration is a meaningful quantity. In such cases the Martin-Löf test based on T given by (3.2) gives a good general impression about the deviations. If T reaches significance, comparison of the signs of the deviations within one item may give valuable clues about the way in which certain items deviate from the expected results. A test based on (3.18) may be a useful formalization of such a sign comparison, which is already provided in the form of plots by some computer programs.

4. Extended Hypergeometric Tests Per Item Pair Per Score Group

4.1 Existing Tests

The assumptions of the Rasch model are monotonicity, sufficiency of total score, unidimensionality, and local independence. The binomial analysis of section 3 is particularly suited for detecting violations of the former two assumptions. A violation of the unidimensionality axiom will usually go hand in hand with a lack of local independence: people with the same total score on a test measuring two different latent traits will not all score the same position on one such trait, and thus items measuring the same trait will be positively related even in such a group. When specific subsets of items are suspected of measuring different latent quantities, one may use a test proposed by Martin-Löf [Note 5], incorporated in the PML program and briefly described by Wainer, Morgan, and Gustafsson [1980]. An interesting test for large sample sizes is suggested by Stene [Note 6].

Van den Wollenberg [1979, 1982] proposes an analysis of local independence via the 2 × 2 tables of pairs of items within score groups. Let N_{rij} denote the number of persons with total score r who have both i and item j correct, and $N_{ri\bar{j}}$ the number in score group r with i correct and j incorrect; similarly $N_{r\bar{i}j}$ and $N_{r\bar{i}\bar{j}}$. Then such a table becomes

observed frequencies item j — probabilities item j

		1	0			1	0	
i	1	N_{rij}	$N_{ri\bar{j}}$	N_{ri}	1	π_{rij}	$\pi_{ri} - \pi_{rij}$	π_{ri}
t								
e								
m	0	$N_{r\bar{i}j}$	$N_{r\bar{i}\bar{j}}$	$N_{r\bar{i}}$	0	$\pi_{rj} - \pi_{rij}$	$1 - \pi_{ri} +$ $-\pi_{rj} + \pi_{rij}$	$1 - \pi_{ri}$
i		N_{rj}	$N_{r\bar{j}}$	n_r		π_{rj}	$1 - \pi_{rj}$	1

Such tables are trivial for $r = 0, 1, k - 1, k$ because of structural zero's; we shall restrict ourselves to $2 \leq r \leq k - 2$ in the rest of this section.

For fixed r, i and j, and given n_r and ε, the unconditional distribution of the entries in the four cells of the 2 × 2 table is multinomial and can be written as

$$P(N_{rij} = x, \quad N_{ri} = n_{ri}, \quad N_{rj} = n_{rj} \mid n_r, \varepsilon)$$

$$= \frac{n_r!\,\pi_{rij}^{x}(\pi_{ri} - \pi_{rij})^{n_{ri}-x}(\pi_{rj} - \pi_{rij})^{n_{rj}-x}(1 - \pi_{ri} - \pi_{rj} + \pi_{rij})^{n_r - n_{ri} - n_{rj} + x}}{x!(n_{ri} - x)!(n_{rj} - x)!(n_r - n_{ri} - n_{rj} + x)!} \tag{4.1}$$

where

$$\begin{aligned} \pi_{rh} &= \frac{\varepsilon_h \gamma_{r-1}^{(h)}}{\gamma_r} \\ \pi_{rgh} &= \frac{\varepsilon_g \varepsilon_h \gamma_{r-2}^{(g,h)}}{\gamma_r}; \end{aligned} \tag{4.2}$$

as before the gamma's denote the symmetric functions and their derivatives.

We shall follow Van den Wollenberg in his preference for a conditional analysis of the 2 × 2 table, assuming N_{ri} and N_{rj} fixed. Possible deviations between N_{ri} and $n_r \pi_{ri}$

already having been detected by the binomial analysis, all we want to do now is find out whether N_{rij} deviates from its conditional expectation given the marginals.

As was already observed by Lehmann [1959, p. 143–146], the conditional distribution of N_{rij} given the marginals is the "extended hypergeometric distribution":

$$P(N_{rij} = x \mid N_{ri} = n_{ri}, \quad N_{rj} = n_{rj}, \quad N_r = n_r, \varepsilon) = C\binom{n_{ri}}{x}\binom{n_r - n_{ri}}{n_{rj} - x}\Delta^x \tag{4.3}$$

where $\Delta = \Delta_{rij}$ is the well known odds ratio

$$\Delta = \frac{\pi_{rij}(1 - \pi_{ri} - \pi_{rj} + \pi_{rij})}{(\pi_{ri} - \pi_{rij})(\pi_{rj} - \pi_{rij})} \tag{4.4}$$

and the factor C is found from the requirement that the sum of (4.3) over its range max $(0, n_{ri} + n_{rj} - n_r) \leq x \leq$ min (n_{ri}, n_{rj}) is equal to 1. Some properties of this extended hypergeometric distribution are given by Patil and Joshi [1968, p. 57]. A full discussion, including moments, estimation methods for Δ and various approximations was given by Harkness [1965]. Rewriting Harkness' theorem 3 in our notation, we obtain the following normal approximation.

Let N_{rij} have the extended hypergeometric distribution given by (4.3). Find the unique numbers p_1 and p_2 for which

$$\begin{aligned} &0 < p_h < 1, \qquad (h = 1, 2) \\ &p_1(1 - p_2) = \Delta\, p_2(1 - p_1) \\ &n_{ri}p_1 + (n_r - n_{ri})p_2 = n_{rj}. \end{aligned} \tag{4.5}$$

Next, let

$$\begin{aligned} H_1 &= \{n_{ri}p_1(1 - p_1)\}^{-1/2} \\ H_2 &= \{(n_r - n_{ri})p_2(1 - p_2)\}^{-1/2} \\ H^2 &= H_1^2 + H_2^2 \\ z_m &= (m - n_{ri}p_1)H. \end{aligned} \tag{4.6}$$

Then $Z = (N_{rij} - n_{ri}p_1)H$ has approximately a standard normal distribution when $n_r \to \infty$ such that $H \to 0$; more specifically, if the integers a and b are chosen in the permissible range given below (4.4) such that H, Hz_a^3 and Hz_b^3 all tend to zero, then the ratio of $P(a \leq N_{rij} \leq b)$ and $\Phi(z_{b+1/2}) - \Phi(z_{a-1/2})$ tends to 1, where Φ denotes the standard normal cdf:

$$\Phi(z) = (2\pi)^{-1/2}\int_{-\infty}^{z} \exp\left(-\tfrac{1}{2}u^2\right) du. \tag{4.7}$$

As (4.3) is an exponential family and has monotone likelihood ratio in Δ, the hypothesis that Δ is given by (4.4) can now be tested either in the exact distribution or in its normal approximation. An extensive numerical comparison has taught us that the normal approximation has less than two-digit accuracy unless $H < \frac{1}{8}$. In most applications, H will be much larger. Fortunately the permissible range will then be moderate enough for exact evaluation with an electronic computer. One could also obtain confidence intervals for Δ using Harkness [1965]. The use of such a diagnostic test per score group per item pair has some advantages compared to the use of an overall test proposed by Van den Wollenberg. One of the obstacles on the road to an overall test is the absence of a fully satisfactory combination procedure for the individual test statistics per item pair and score group, which have different discrete distributions and are slightly dependent because of some side conditions analogous to (3.8), (3.9), and (3.11).

Before proceeding with this discussion, let us compare our results to those obtained by Van den Wollenberg [1982], who claims in his formula (20) that

$$E(N_{rij} \mid n_r, \boldsymbol{\varepsilon}) = n_r \pi_{rij} = \frac{n_r \varepsilon_i \varepsilon_j \gamma_{(r-2)}^{(i,j)}}{\gamma_r}. \tag{4.8}$$

He next proposes to compare observed and expected frequencies by means of

$$Q_{rij} = \frac{D^2}{E(N_{rij})} + \frac{D^2}{E(N_{r\bar{i}j})} + \frac{D^2}{E(N_{ri\bar{j}})} + \frac{D^2}{E(N_{r\bar{i}\bar{j}})}, \tag{4.9}$$

where D denotes the difference between N_{rij} and its expectation. Finally he claims that Q_{rij} has a chi-square distribution with one degree of freedom, and combines them into

$$Q_{2(r)} = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} Q_{rij} \frac{(k-3)}{(k-1)}. \tag{4.10}$$

which is claimed to have approximately a chi-square distribution with $\frac{1}{2}k(k-3)$ degrees of freedom.

Examining the appropriateness of the three steps we shall distinguish four cases.

Case 1: conditioning on n_r and using general values $\boldsymbol{\varepsilon}$.

Here the distribution of the four entries of the table is the multinomial (4.1) and $E(N_{rij})$ is correctly given by (4.8). Next,

$$D_{\bar{i}j} = N_{r\bar{i}j} - EN_{r\bar{i}j} = N_{rj} - N_{rij} - n_r(\pi_{rj} - \pi_{rij}) = N_{rj} - n_r \pi_{rj} - D, \tag{4.11}$$

and similarly for the remaining two cells. The four squared deviations thus are not identical; indeed the table has three degrees of freedom and only the fourth entry is fixed by the remaining three through the requirement that they sum to n_r. Some tedious algebra shows that Q_{rij} given by (4.9) equals $cD^2/\{n_r \pi_{rij}(1 - \pi_{rij})\}$: divided by c it would be asymptotically ($n_r \to \infty$) chi-square with one degree of freedom, but the (complicated) expression c is a constant which will not equal 1 unless some very special choice is made for $\boldsymbol{\varepsilon}$.

Case 2: conditioning on n_r and substituting for $\boldsymbol{\varepsilon}$ the values estimated by conditional maximum likelihood for this score group alone.

Again $E(N_{rij})$ is given by (4.8), with the estimated ε inserted. As observed by Van den Wollenberg, the special estimates $\hat{\boldsymbol{\varepsilon}}_r = (\hat{\varepsilon}_{r1}, \hat{\varepsilon}_{r2}, \ldots, \hat{\varepsilon}_{rk})$ are such that

$$\hat{\pi}_{ri} \overset{\text{def}}{=} \frac{\hat{\varepsilon}_{ri} \gamma_{r-1}^{(i)}(\hat{\boldsymbol{\varepsilon}}_r)}{\gamma_r(\hat{\boldsymbol{\varepsilon}}_r)} = \frac{N_{ri}}{n_r}. \tag{4.12}$$

It follows that in (4.11) now $D_{ij} = -D$; similarly $D_{i\bar{j}} = -D$ and $D_{i\bar{j}} = D$. If we put

$$\begin{aligned} p_1 &= \frac{\hat{\pi}_{rij}}{\hat{\pi}_{ri}} = \frac{\hat{\pi}_{rij} n_r}{N_{ri}}, \\ p_2 &= \frac{(\hat{\pi}_{rj} - \hat{\pi}_{rij})}{(1 - \hat{\pi}_{ri})} = \frac{(N_{rj} - n_r \hat{\pi}_{rij})}{(n_r - N_{ri})}, \end{aligned} \tag{4.13}$$

then, taking the first and third, and also the second and fourth term together, some algebra shows that

$$\frac{1}{E(N_{rij})} + \frac{1}{E(N_{r\bar{i}j})} + \frac{1}{E(N_{ri\bar{j}})} + \frac{1}{E(N_{r\bar{i}\bar{j}})} = \frac{1}{N_{ri} p_1(1 - p_1)} + \frac{1}{(n_r - N_{ri}) p_2(1 - p_2)}, \tag{4.14}$$

Table 3. A summary of properties of the four cases.

	Case 1	Case 2	Case 3	Case 4
Conditioning on $\underset{\sim}{\varepsilon}$ and n_r	yes	yes	yes	yes
Conditioning on n_{ri} and n_{rj}	no	no	yes	yes
General values of $\underset{\sim}{\varepsilon}$	yes	no	yes	no
$\underset{\sim}{\varepsilon}$ by CML for score group r	no	yes	no	yes
$EN_{rij} = n_r \pi_{rij}$	yes	yes	no	yes
$D_{ij} = -D_{i\bar{j}} = -D_{\bar{i}j} = D_{\bar{i}\bar{j}}$	no	yes	yes	yes
Q_{rij} has χ_1^2 distribution	no	no	no	asympt.

a result that we shall need in case 4. For a proof of the asymptotic chi-square character of (4.9) we should prove that (4.14) asymptotically equals $\{n_z \hat{\pi}_{zij}(1 - \hat{\pi}_{zij})\}^{-1}$; we shall obtain in case 4 that this is not correct.

Case 3: conditioning on n_r, n_{ri}, n_{rj} and using general values of ε.

The distribution of N_{rij} now is extended hypergeometric, see (4.3). The mean $E(N_{rij})$ is no longer given by (4.8). No simple closed expression for it exists [Harkness, 1965], but it will be clear that $E(N_{rij})$ will be larger than $n_r \pi_{rij}$ when $N_{ri} > n_r \pi_{ri}$ and $N_{rj} > n_r \pi_{rj}$, and smaller when the reverse inequalities hold.

By the conditioning on the marginals, the four squared deviations from the expectations are trivially equal. As an asymptotic chi-square distribution will be obtained in case 4 for a very special vector ε, it is extremely plausible that this will not hold here, with the use of a wrong expectation and of a variance which is not $1/H^2$ (see below).

Case 4: conditioning on n_r, n_{ri}, n_{rj} and substituting for ε *the values estimated by conditional maximum likelihood for this score group alone.*

As in case 3, the expectation of N_{rij} cannot be obtained in closed form, the distribution being again extended hypergeometric. But the numbers p_1 and p_2 defined by (4.13) now satisfy the three requirements (4.5) by Harkness, as is easily verified. It follows that $n_r \hat{\pi}_{rij} = n_{ri} p_1$ is equal to the asymptotic mean used in Harkness' normal approximation given below (4.6); thus $E(N_{rij})$ and $n_r \hat{\pi}_{rij}$ are asymptotically equal. The equality of the four squared deviations follows from the conditioning, as in case 3. For the asymptotic distribution of Q_{rij}, note that (4.14) can now be written as H^2 by (4.6), and thus Q_{rij} is simply the square of a quantity of which the asymptotic normality was proven by Harkness.

The four cases are summarized in Table 3. We conclude that the claim by Van den Wollenberg is asymptotically correct in case 4; in all other cases one of its parts fails to be true.

4.2 New Proposals

Earlier we have complained that most test proposals were global, asymptotic, and concentrated on the null hypothesis. This is very true for $Q_{2(r)}$ given in (4.10): it does not consider separate items or item pairs, its components Q_{rij} are based on 2 × 2 tables with rather small entries unless all score groups contain many persons, and it is powerful against all possible deviations between the quantities N_{rij} and their expectations under the Rasch model.

One could argue that the distribution of $Q_{2(r)}$, being the sum of $\binom{k}{2}$ quantities Q_{rij}, will be almost chi-square under H_0 even for moderate sample sizes, because so many contributions are summed. On closer inspection, this argument is not very compelling. Unless one has a large sample size and simultaneously a small number of items, a substantial number of score groups may contain ten or less individuals. By the restrictions specified below (4.4), this means that N_{rij} can assume at most a few different values; for some item pairs only two or even only one value may be possible under the conditioning on N_{ri} and N_{rj}. For such cases Q_{rij} will have a strange discrete distribution under H_0 and little or no power against general violations of the Rasch model.

In order to improve this by a directed search, let us analyze in more detail the null hypothesis that items i and j belong to the same Rasch scale. People in the same score group obtain the same estimated ability. At first sight this seems to imply that local independence is equivalent to $\Delta = 1$, that is to

$$E(N_{rij} \mid n_{ri}, n_{rj}, n_r, \varepsilon) = \frac{n_{ri} n_{rj}}{n_r} \tag{4.15}$$

in the conditional distribution (4.3) of N_{rij}. But within the score group we have brought together individuals of somewhat different latent trait value, although with the same estimated ability. As they have the same number r of items correct, individuals with item i correct will have a slightly lower probability of having item j correct than the average person in score group r. Even if the Rasch model is correct, the values of Δ given by evaluating (4.4) with (4.2) calculated for a vector of estimates $\hat{\varepsilon}$ will therefore generally be slightly less than one: *within score group r, items i and j will have a weak negative correlation.* The size of this correlation will be larger when the scale contains fewer items: for large k the restriction to total score r will have almost no influence.

Let us now turn to the specific alternative in which the item set consists of some items measuring one latent trait and some other items measuring a different latent trait. The numbers correct on the two subsets of items, within the total score r, can then be any pair

$$(r, 0), \quad (r-1, 1), \quad (r-2, 2), \ldots, (1, r-1), \quad (0, r), \tag{4.16}$$

subject to the restriction that some pairs may be impossible because the number of items per subtrait is limited.

Table 4 sketches the situation in which a test of $k = 10$ items consists of 6 items measuring one trait and 4 items measuring a different one. Especially within the groups having a total score of 4, 5, or 6, people have widely differing scores on each of the subtests.

In such a heterogeneous score group, we predict that two items from the same subtest will be positively correlated. A pair consisting of one item from each subtest should

Table 4. Possible combinations of scores on two subsets consisting of 6 and 4 items respectively; in each cell the total score is printed.

score on	4	4	5	6	7	8	9	10	
subtest 2	3	3	4	5	6	7	8	9	
	2	2	3	4	5	6	7	8	
	1	1	2	3	4	5	6	7	
	0	0	1	2	3	4	5	6	
		0	1	2	3	4	5	6	score on subtest 1

show almost independence when both latent traits are uncorrelated. When the total number of items is small, the side effect mentioned under the null hypothesis also plays a role. It may then be expected, however, that item pairs from the same subset have a higher Δ than predicted by the Rasch model and item pairs from different subsets have a lower Δ.

This result may be summarized as follows. For each score r $(2 \leq r \leq k-2)$ and each item pair (i, j), compare

$$\bar{\Delta}_{rij} = \frac{N_{rij}(n_r - n_{ri} - n_{rj} + N_{rij})}{(n_{ri} - N_{rij})(n_{rj} - N_{rij})} \tag{4.17}$$

to Δ_{rij} given by inserting the estimates of ε in score group r into (4.2) and (4.4). If the item set can be subdivided into two subsets such that generally $\bar{\Delta}_{rij} > \Delta_{rij}$ for item pairs from the same subset and $\bar{\Delta}_{rij} < \Delta_{rij}$ from item pairs consisting of one from each subset, then this is evidence that each subset of items measures a different latent trait. The significance of the difference between $\bar{\Delta}_{rij}$ and Δ_{rij} can be tested using the extended hypergeometric distribution (4.3). If all cell entries are large, one may also use the Harkness approximation, which comes close to calculation of Q_{rij} by (4.9) and (4.8) when estimates of ε per score group are used. In most cases it is better to use the one-sided version with continuity correction, and Φ defined in (4.7);

$$\begin{aligned} P(N_{rij} \leq x) &\approx \Phi[(x + \tfrac{1}{2} - n_r \hat{\pi}_{rij})H], \\ P(N_{rij} \geq x) &\approx 1 - \Phi[(x - \tfrac{1}{2} - n_r \hat{\pi}_{rij})H], \end{aligned} \tag{4.18}$$

where

$$H^2 = \frac{1}{\hat{\pi}_{rij}} + \frac{1}{\hat{\pi}_{ri} - \hat{\pi}_{rij}} + \frac{1}{\hat{\pi}_{rj} - \hat{\pi}_{rij}} + \frac{1}{1 - \hat{\pi}_{ri} - \hat{\pi}_{rj} + \hat{\pi}_{rij}} \tag{4.19}$$

and all $\hat{\pi}$ values are obtained by inserting the estimates of ε for this score group alone into (4.2).

More generally, we predict that item pairs for which $\bar{\Delta}_{rij} > \Delta_{rij}$ for (almost) all score groups, especially for all score groups with large n_r, point to a very close association between such items, that might measure a subtrait which is less well measured by at least some of the remaining items. Combination of one-sided tests for item pairs could be developed in a way analogous to our derivation of (3.18).

Once more we shall use the items on pupils' attitudes for an illustration. We shall use two different scales of six items, each of which consists of three items from the first subscale, and three items from the second one. Our selection was based on the two requirements that (a) score groups 2, 3, and 4 should contain at least 70 persons each and (b) the three items chosen from the same subscale should not exhibit clear departures from the Rasch model.

Table 5 gives the results. In each half of the table the first six lines give $\bar{\Delta}_{ij}$ and Δ_{ij} for the pairs within a subscale, and the next nine lines for the pairs consisting of one item from each subscale. Thus we expect $\bar{\Delta}_{ij} > \Delta_{ij}$ in the first six lines and $\bar{\Delta}_{ij} < \Delta_{ij}$ in the next nine lines. Such inequalities indeed hold in 78 out of the 90 cases listed in the table; the twelve offenders are marked by printing a star instead of the expected sign.

Note that, for $k = 6$ items only, all expected values Δ_{rij} are far less than one: by the restriction to a fixed total score r all pairs of items exhibit negative association. Observed values for item pairs within the same subscale are sometimes larger than one and sometimes smaller.

Table 5. Observed $\bar{\Delta}_{rij}$ calculated from (4.17) and expected Δ_{rij} from the $\underset{\sim}{\varepsilon}$ estimates per score group inserted into (4.2) and (4.4). Inequalities violating our prediction are replaced by a star. Items 2,7,10,17,18,19(first half) and 6,9,10,15,16,18 (second half).

item pair		r=2 , n=70			r=3 , n=105			r=4 , n=128		
i	j	$\bar{\Delta}_{rij}$		Δ_{rij}	$\bar{\Delta}_{rij}$		Δ_{rij}	$\bar{\Delta}_{rij}$		Δ_{rij}
2	7	2.73	>	0.36	1.01	>	0.44	0.17	*	0.35
2	10	0.62	>	0.31	1.56	>	0.39	0.52	>	0.34
7	10	0.63	>	0.34	1.00	>	0.42	8.66	>	0.35
17	18	1.28	>	0.33	0.94	>	0.40	0.96	>	0.32
17	19	1.38	>	0.31	1.26	>	0.39	1.03	>	0.32
18	19	0.59	>	0.32	0.76	>	0.38	0.16	*	0.33
2	17	0.04	<	0.30	0.18	<	0.40	0.19	<	0.33
2	18	0.06	<	0.30	0.44	*	0.40	0.43	*	0.34
2	19	0.06	<	0.27	0.09	<	0.39	0.39	*	0.34
7	17	0.05	<	0.33	0.32	<	0.43	0.04	<	0.34
7	18	0.35	*	0.34	0.09	<	0.42	0	<	0.35
7	19	0.25	<	0.32	0.44	*	0.42	0	<	0.35
10	17	1.71	*	0.34	0.10	<	0.40	0.12	<	0.36
10	18	0	<	0.34	0.21	<	0.39	0.21	<	0.34
10	19	0.43	*	0.32	0.20	<	0.37	0.20	<	0.33
		r=2 , n=79			r=3 , n=105			r=4 , n=110		
6	9	0.93	>	0.25	0.64	>	0.35	0.52	>	0.29
6	10	0.98	>	0.28	2.69	>	0.32	1.02	>	0.23
9	10	0.48	>	0.29	2.50	>	0.37	4.46	>	0.26
15	16	0.62	>	0.23	1.63	>	0.37	0.54	>	0.34
15	18	2.11	>	0.29	0.71	>	0.31	1.38	>	0.34
16	18	0.58	>	0.19	3.38	>	0.35	0.48	>	0.24
6	15	0	<	0.29	0.22	<	0.32	0.11	<	0.34
6	16	0.03	<	0.18	0.09	<	0.34	0.29	*	0.26
6	18	0.15	<	0.27	0.07	<	0.24	0.10	<	0.23
9	15	0.05	<	0.30	0.05	<	0.38	0.04	<	0.36
9	16	0.47	*	0.35	0.52	*	0.41	0	<	0.29
9	18	0.07	<	0.26	0.22	<	0.36	0	<	0.27
10	15	0	<	0.31	0.28	<	0.36	0.28	<	0.34
10	16	0.07	<	0.22	0.02	<	0.37	0	<	0.24
10	18	0	<	0.29	0.06	<	0.31	0.02	<	0.20

5. *Discussion*

The tools proposed in this paper could be viewed as a specialization, for the case of dichotomous and equi-discriminating items, of more general principles applicable in other measurement models. The splitter item technique plots the estimated item difficulties $\hat{\sigma}_j$ for all items $j \neq i$ in the subgroups with item i correct and wrong; if N_{ij} persons score both items right, then such estimates are based on the sufficient statistics N_{ij} and $N_i - N_{ij}$ respectively. Essentially the association between items i and j is plotted. Within the Rasch model, however, the transformation from observed dichotomous scores to population free item estimates allows an unbiased form of such plots.

Similarly, the binomial analysis in section 3 can be viewed as a refined item-test correlation and the extended hypergeometric analysis as a benchmark for a simple cluster analysis based on the odds ratio Δ. In both cases the conditioning on total score is similar to elimination of the first principal component. The difference is again that our analysis is specifically geared to the special assumptions of the Rasch model: it is more complicated than a classical form of item analysis, because it is carried out within a specifically objective measurement model for dichotomous items.

The results in sections 3 and 4 are special cases of a comparison of observed and expected frequencies for all score patterns, cf. Table 8 column 6 of Andersen [1982]. One could say that we have grouped such frequencies according to the result for one item, or for a pair of items. Low frequencies will presumably make a grouping desirable, and our grouping has the advantage of pinpointing properties of specific items. A search for specific violations in the spirit of our paper, but now based on asymptotic properties, was advocated by Kelderman [Note 7].

The Birnbaum model with unequal item discriminations comes very close to the Rasch model when the slopes are assumed known [Andersen, Note 8]. Due to the high correlation between various weighted composites of item scores, the use of somewhat wrong slopes may not have serious effects. Lord [1981] advocates the estimation of both slopes and guessing parameters jointly with item difficulties. Our proposals are not directly relevant for this case: the applications we have in mind do involve neither multiple choice achievement tests nor very large sample sizes.

An overall goodness of fit test frequently leads to the conclusion that a given data matrix of scores of n persons on k dichotomous items does not obey the Rasch model. The exploratory tools proposed in this paper may be helpful in transforming this general conclusion to more specific information about the nature of the violations. They deal with most of the themes outlined in section 1. Although the violations may demand the use of more general item response models mentioned above, it may also be desirable to investigate whether removal of a few items brings us within the Rasch model with its desirable statistical properties.

We agree with Gustafsson [1980] that some violations are not caused by the item itself but by the conditions under which it is taken: removal of items should never become a mechanical operation in which item content and circumstances are disregarded. Similar tools could be developed which exhibit outlying persons or subgroups of "Rasch homogeneous persons", rather than items. Finally, "significant" violations of the Rasch assumptions may be tolerable when they are only detected because a very large sample size has made the tests extremely powerful. More work on the robustness of procedures based on the Rasch model against minor violations of the assumptions would be highly welcome.

REFERENCE NOTES

1. Allerup, P. & Sorber, G. *The Rasch model for questionnaires. With a computer program* (2nd ed). Copenhagen: The Danish Institute for Educational Research, 1977, 4.
2. Gustafsson, J. E. The Rasch model for dichotomous items: Theory applications and a computer program. *Reports from the Institute of Education*, University of Göteborg, no. 63, 1977.
3. Gustafsson, J. E. PML: A computer program for conditional estimation and testing in the Rasch model for dichotomous items, *Reports from the Institute of Education*, University of Göteborg, no. 85, 1979.
4. Raaijmakers, N. H. & Van den Wollenberg, A. L. Radi: program for the dichotomous Rasch model. *Program Bulletin*, University of Nijmegen, 1979.
5. Martin-Löf, P. *Statistika modeller.* Anteckningar från seminarier läsaret 1969–70 utarbetade av Rolf Sundberg, 2 : a uppl. (Statistical models. Notes from seminars 1969–70 by Rolf Sundberg, 2nd ed.) Institute för försäkringsmatematik och matematisk statistik vid Stockholms universitet, 1973.
6. Stene, E. *An exact test for stochastic independence of responses in an item analysis model*, Symposium on Rasch models, Køge (Denmark), 1969.

7. Kelderman, H. (1982) *Loglinear Rasch model tests.* Submitted to Psychometrika.
8. Andersen, E. B. *A general latent structure model for contingency table data,* paper presented in Princeton, May 1982, to be published.

REFERENCES

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

Andersen, E. B. Latent trait models and ability parameter estimation. *Applied Psychological Measurement,* 1982, to be published.

Fischer, G. H. *Einführung in die Theorie Psychologischer Tests. Grundlagen und Anwendungen.* Bern: Huber, 1974.

Fischer, G. H. & Scheiblechner, H. Algorithmen und Programma für das probabilistische Testmodell von Rasch. *Psychologische Beiträge,* 1970, *12,* 23–51.

Formann, A. K. Über die Verwendung von Items als Teilungskriterium für Modellkontrollen im Modell von Rasch, *Zeitschrift für Experimentelle und Angewandte Psychologie,* 1981, *28,* 541–560.

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

Harkness, W. L. Properties of the Extended Hypergeometric distribution, *Annals of Mathematical Statistics,* 1965, *36,* 938–945.

Lehmann, E. L. *Testing Statistical Hypotheses,* Wiley: New York, 1959.

Lord, F. M. *Applications of item response theory to practical testing.* Erlbaum, Hillsdale N.J., 1981.

Patil, G. G. & Joshi, S. W. *A dictionary and bibliography of discrete distributions.* Oliver & Boyd: Edinburgh, 1968.

Van den Wollenberg, A. L. *The Rasch model and time-limit tests,* Ph.D. thesis, University of Nijmegen, 1979.

Van den Wollenberg, A. L. Two new test statistics for the Rasch model, *Psychometrika,* 1982, *47,* 123–140.

Wainer, H., Morgan, A. & Gustafsson, J. E., A review of estimation procedures for the Rasch model with an eye toward longish tests, *Journal of Educational Statistics,* 1980, *5,* 35–64.

Wallis, W. A. Compounding probabilities from independent significance tests, *Econometrica,* 1942, *10,* 229–248.

Wright, B. D. & Panchapakesan, N. A procedure for sample-free item analysis. *Educational and Psychological Measurement,* 1969, *28,* 23–48.

Wright, B. D. & Stone, M. H. *Best test design, Rasch measurement,* MESA PRESS, 5835 Kimbark Ave. Chicago (Ill), 1979.

Manuscript received 2/23/81
Final version received 7/8/82