Open access • Journal Article • DOI:10.1348/000711001159401

Some Mantel-Haenszel tests of Rasch model assumptions. — Source link

Tom Verguts, Paul De Boeck

Institutions: Katholieke Universiteit Leuven

Published on: 01 May 2001 - British Journal of Mathematical and Statistical Psychology (Blackwell Publishing Ltd)

Topics: Polytomous Rasch model, Rasch model and Cochran–Mantel–Haenszel statistics

British Journal of Mathematical and Statistical Psychology (2001), **54**, 21–37 *Printed in Great Britain*

Some Mantel–Haenszel tests of Rasch model assumptions

Tom Verguts* and Paul De Boeck

University of Leuven, Belgium

A class of Rasch model tests is proposed, all of them based on the Mantel–Haenszel chi-squared statistic. All tests make use of the 'sufficient statistics' property the Rasch model possesses. One element of our general class, the test for item bias developed by Holland and Thayer, has been discussed extensively in the psychometric literature. Three applications of the general procedure are presented, two on unidimensionality and one on item dependence in educational testing. In each case, simulation results are reported. Our procedure is also applied to real data.

1. Introduction

Suppose X_{pi} is a dichotomous random variable which indicates the score (1 or 0) for person p on item i ($p = 1, \ldots, P$; $i = 1, \ldots, I$). With person p we associate an ability parameter ξ_p and with item i a difficulty ϵ_i, both parameters positive-valued. The Rasch model can then be stated as

$$\Pr(X_{pi} = 1 | \xi_p, \epsilon_i) = \frac{\xi_p \epsilon_i}{1 + \xi_p \epsilon_i}. \tag{1}$$

It can be shown that model (1) is equivalent with the following five assumptions (Fischer, 1995a): (i) unidimensionality of the 'latent trait' ξ; (ii) monotonicity of $\Pr(\cdot)$ in ξ; (iii) lower limit 0 and upper limit 1 of $\Pr(\cdot)$ for ξ going to 0 and $+\infty$, respectively (denoted *no guessing*, for short). The two key assumptions, which will be used in the following, are: (iv) local stochastic independence; and (v) sufficiency of raw sum score for ξ_p and number correct per item for ϵ_i. (Actually, these five assumptions are equivalent to a 'family of Rasch models', in the sense that the factor $\xi_p \epsilon_i$ in (1) may be replaced with $b(\xi_p \epsilon_i)^a$ for any choice of a, b where both $a, b > 0$. However, without loss of generality we choose the constants $a = b = 1$; see Fischer, 1995a.) In the following, the phrase 'under the Rasch model' will denote the occurrence of these five assumptions together. Sufficiency is the key property used here and will be elaborated later on.

The methodology developed in this paper is a contingency table approach to testing Rasch model assumptions. Rather than estimating item or person parameters and then performing goodness-of-fit tests, this method directly uses contingency table(s) constructed from the data to test model assumptions. We base our test on a fundamental property of the Rasch model,

* Requests for reprints should be addressed to Tom Verguts, Department of Psychology, Katholieke Universiteit Leuven, Tiensestraat 102, B-3000 Leuven, Belgium (e-mail: Tom.Verguts@psy.kuleuven.ac.be).

namely, that the conditional distribution of the responses given the total raw score is independent of the person parameters. This property forms the basis, for example, for conditional maximum likelihood estimation (Fischer, 1974) and the conditional likelihood ratio test (Andersen, 1973).

One procedure that uses this property was developed by Holland & Thayer (1988). These authors use the fact that, under the Rasch model, gender (or any other binary, external, criterion) and 'score on item i' should be (statistically) independent variables given a certain score level t (which follows from the fundamental property referred to in the previous paragraph). The item i should be included in the calculation of t; we will come back to this point later. Since gender and 'score on item i' are independent, the Mantel–Haenszel (MH) chi-squared statistic (Mantel & Haenszel, 1959) based on these variables is distributed as chi-squared with one degree of freedom. Further studies of this method can be found in Parshal & Miller (1995), Uttaro & Millsap (1994), Zwick (1990), and Zwick, Donoghue & Grima (1993).

In this paper we extend the differential item functioning (DIF) application by showing that the MH methodology can be used to test, besides DIF, other Rasch model assumptions as well. The key point is an appropriate choice of row and column headings in the MH table. For appropriately constructed tables, the Rasch model predicts that the MH chi-squared statistic should be χ^2 distributed. On the other hand, if the MH chi-squared statistic turns out to be too high, this indicates that some aspect of the model is violated.

Many testing procedures for the Rasch model have been described in the literature. We distinguish two such types of procedures, the first parametric, in which (estimated) person and/or item parameters are used, and the second non-parametric, which involves constructing a test without the need for these parameters. The approach adhered to in this paper is the non-parametric one.

One class of commonly used parametric tests is the set of likelihood ratio tests. An example is a test proposed by Martin-Löf—see Glas & Verhelst (1995), whose terminology we follow, and Gustafsson (1980). The key idea here is to estimate the item parameters ϵ_i in two separate groups of items and also in the total group of items, and then to check (with a likelihood ratio test) whether the two sets of estimated parameters conform to each other. Let the items be partitioned in two sets consisting of I_1 and I_2 items respectively. Let $\mathbf{t} = (t_1, t_2)$ denote a vector of scores on the first and the second set of items respectively, and $\mathrm{n}_{\mathbf{t}}$ the number of people with this score pattern; n_t is the number of people with a score t on the total item set. The Martin-Löf statistic is then defined as

$$LR = 2\left(\sum_{\mathbf{t}} n_{\mathbf{t}} \ln\left(\frac{n_{\mathbf{t}}}{P}\right) - \sum_{t} n_t \ln\left(\frac{n_t}{P}\right) - \ln L_C + \ln L_C^{(1)} + \ln L_C^{(2)}\right), \tag{2}$$

in which L_C, $L_C^{(1)}$, and $L_C^{(2)}$ denote the likelihood functions based on the total, the first and the second item set respectively, evaluated in the conditional maximum likelihood estimators. Under the Rasch model, this statistic has a chi-squared distribution with $I_1 I_2 - 1$ degrees of freedom. The items can be assigned to the two sets in different ways, yielding different tests of the model (Gustafsson, 1980). For example, if the items are grouped according to difficulty, the test yields a test of differing person slopes. If the items are grouped according to two purported underlying dimensions, the statistic tests unidimensionality between the two sets of items. This second application will be discussed later.

Concerning the non-parametric approach, we have already mentioned the Holland and Thayer (1988) DIF application. A second non-parametric approach to test model assumptions was developed by Rosenbaum (1984), although the tests were derived under some general assumptions (monotonicity and conditional independence), rather than under the Rasch model (which entails a few more assumptions). In this procedure, participants are ordered using a subset J_1 of the test. At each level of the score on J_1, the scores on the remaining part of the test should be associated, meaning that any increasing functions $g_1(J_2)$ and $g_2(J_2)$ should be non-negatively correlated. For example, if participants are grouped according to their score on all items except items i and j, these items should be correlated at each level of this subscore. Otherwise, the monotonicity assumption would be violated (for these two items). Rosenbaum constructs a test to investigate this correlation. It is interesting to note that this author also uses the MH chi-squared statistic and that some of his methods are similar (but not identical) to the procedures developed here.

Another non-parametric procedure was developed by Ponocny (1999; Ponocny & Ponocny-Seliger, 1999), extending an idea presented by Rasch (1966), namely that under the Rasch model every data matrix with the same marginals has the same probability of occurring. From this property, Ponocny shows how a uniformly most powerful test of the Rasch model against a large set of alternative hypotheses may be constructed. This entails calculating a statistic T in the observed data set and in all other matrices with the same marginal totals (i.e., person sum scores and item scores), and checking the proportion of matrices where T is larger than the value of T in the observed data matrix; this gives the desired p-value of the Rasch model versus the alternative. A problem with this procedure is the phrase 'all other matrices with the same marginals'; it turns out to be very difficult to enumerate all these matrices. Ponocny presents an ingenious algorithm to obtain or approximate the desired proportion, although in practice the procedure is still limited to data matrices of moderate size.

A first (practical) advantage of using the MH chi-squared statistic to test model assumptions is its ease of use. No parameters need to be estimated, and the MH chi-squared statistic is part of many standard statistical software packages. Also, the computation time is very low, even for large data matrices. A second advantage is that they are all quite specific; that is, they are sensitive to specific types of model violation. The corresponding drawback is that if the MH table is incorrectly specified, then the model will not be rejected even if there are model misspecifications.

The remainder of this paper is organized as follows. First, we introduce the concept of sufficiency and discuss some ways to test the Rasch model based on this concept. Then, we present our general class of Rasch model tests, followed by three applications and corresponding simulation studies. Finally, we present an analysis of a real data set with our methodology.

2. Sufficiency, independence, and the MH chi-squared statistic

The fact that model (1) has sufficient statistics for its parameters follows from being a member of the exponential family (Mood, Graybill & Boes, 1974). For the ξ_p parameter, the sufficient statistic is the raw sum score $T_p \equiv \Sigma_i X_{pi}$, with realizations t_p. Sufficiency means that

$$\Pr(X_{pi} = 1 | t_p, \xi_p, \boldsymbol{\epsilon}) = \Pr(X_{pi} = 1 | t_p, \boldsymbol{\epsilon}), \tag{3}$$

Table 1. Item bias data

	X_i	
G	1	0
1	n_{11t}	n_{10t}
0	n_{01t}	n_{00t}

where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_I)$. That is, all persons within a score group have the same probability of solving an item correctly.

Suppose every person is assigned to one of two groups based on a criterion external to the test, for example, male or female. The variable G will indicate this, so $G = 0$ can stand for male, $G = 1$ for female. It is clear from the sufficiency property (3) that under the Rasch model, the factors 'membership' (G) and 'score on item i' (1 or 0) are independent, given the score level t. Formally, the Rasch model implies

$$\Pr(X_{pi} = 1 | t_p, G = 1) = \Pr(X_{pi} = 1 | t_p, G = 0). \tag{4}$$

Note that the matching variable (in this case t) includes the studied item i (Holland & Thayer, 1988).

From this result, Holland and Thayer (1988) developed a test for DIF starting from success and failure counts for a given item in two groups, as represented in Table 1, for every score group t. The row headings indicate group membership ($G = 0$ or 1, e.g., male or female), and column headings denote the score on item i. In the upper left cell, for example, we have n_{11t}, the number of persons who are member of group $G = 1$, scored item i correctly ($= 1$), and achieved a score of (exactly) t. Then, one may construct the Mantel–Haenszel statistic MH, which is defined as

$$MH = \frac{\left(\sum_t N_{11t} - \sum_t \mathrm{E}(N_{11t}|n_t)\right)^2}{\mathrm{Var}\left(\sum_t N_{11t} | n_1, n_2, \ldots, n_{\mathrm{I}-1}\right)} = \frac{\left(\sum_t N_{11t} - \sum_t \mathrm{E}(N_{11t}|n_t)\right)^2}{\sum_t \mathrm{Var}(N_{11t}|n_t)}, \tag{5}$$

with $E(\cdot)$ and $Var(\cdot)$ denoting mean and variance respectively, where

$$E(N_{11t}|n_t) = (N_{11t} + N_{01t})(N_{11t} + N_{10t})/n_t,$$

$$Var(N_{11t}|n_t) = \mathrm{E}(N_{11t}|n_t)[(N_{01t} + N_{00t})/n_t][(N_{10t} + N_{00t})/(n_t - 1)],$$

$$n_t = N_{11t} + N_{10t} + N_{01t} + N_{00t}$$

and n_t equals the number of subjects who belong to score-group t. Each summation is taken over $t = 1, \ldots, I - 1$, since for $t = 0$ or I the contribution is always zero. It is possible to correct for continuity in (5), but this will not be pursued here. Under independence of G and X_i, the MH chi-squared statistic is asymptotically χ^2 distributed with one degree of freedom, so if the Rasch model holds, MH should be asymptotically χ^2 distributed. On the other hand, high values of MH are indicative of DIF in item i.

Fischer (1993, 1995b) generalizes this result in the following way. He proposes

considering as column headings not responses to single items but rather to two items, so that, for items (i,j), the possible patterns are (1, 1), (1, 0), (0, 1) and (0, 0). He shows that the Rasch model predicts independence in the resulting 2×4 table and proposes a Pearson χ^2 statistic per table. This statistic has $(2-1)(4-1)$ degrees of freedom per table. Since the values can be added over all score groups $t = 2, \ldots, I-2$, the resulting statistic has $3(I-3)$ degrees of freedom. This extended procedure has the advantage that it is also sensitive to model violations caused by item interactions (i.e., association between items).

We will generalize the Holland and Thayer (1988) idea in two ways. First, we show that other criteria for choosing the column headings (rather than $X_i = 0, 1$) can be employed. Second, we show that choosing the row headings can also be based on an *internal criterion*, that is a criterion based on the observed responses, rather than just on an external criterion such as male/female. This allows us to use the MH chi-squared statistic to test other model assumptions besides item bias and item interactions.

3. A general class of MH Rasch model tests

Let the complete set of I items be partitioned in two sets of I^{R} and I^{C} items respectively. Note that $I^{\mathrm{R}} + I^{\mathrm{C}} = I$. The two sets of items will be denoted J^{R} and J^{C}, consisting of I^{R} and I^{C} items respectively. (The R/C notation, for 'rows' and 'columns', will become clear later on.) The score on the item sets J^{R} and J^{C} is denoted by t_p^{R} and t_p^{C}, respectively (so $t_p = t_p^{\mathrm{R}} + t_p^{\mathrm{C}}$). Response patterns are denoted by $\mathbf{w} = (w_1, \ldots, w_I)$, and they can analogously be partitioned as $\mathbf{w} = (\mathbf{w}^{\mathrm{R}}, \mathbf{w}^{\mathrm{C}})$, with corresponding random variable $\mathbf{W}^{\mathrm{C}}$ for $\mathbf{w}^{\mathrm{C}}$.

For example, all response patterns $\mathbf{w}^{\mathrm{C}}$ are based on the item set J^{C}. Consider a level of the score variable t^{C}, and consider all patterns $\mathbf{w}^{\mathrm{C}}$ that result in this score (i.e., all response patterns $\mathbf{w}^{\mathrm{C}}$ with $\Sigma_i w_i^{\mathrm{C}} = t^{\mathrm{C}}$). Denote this set by $\Omega(t^{\mathrm{C}})$; it consists of $\binom{I^{\mathrm{C}}}{t^{\mathrm{C}}}$ elements.

To illustrate these definitions, consider a test of $I = 4$ items, and $J^{\mathrm{C}} = \{$item 1, item 2, item 3$\}$. Further, consider $t^{\mathrm{C}} = 1$, hence $\Omega(t^{\mathrm{C}}) = \{(1,0,0), (0,1,0), (0,0,1)\}$, which are $\binom{3}{1} = 3$ elements. Take an arbitrary proper subset of this set, and denote it by $\Omega(t^{\mathrm{C}}, 1)$. The corresponding set of remaining response patterns is denoted $\Omega(t^{\mathrm{C}}, 2)$. In the example, $\Omega(t^{\mathrm{C}}, 1)$ might be equal to $\{(1,0,0)\}$ and hence $\Omega(t^{\mathrm{C}}, 2) = \{(0,1,0), (0,0,1)\}$.

This terminology allows one to construct MH tables in the following way. First, classify every person according to her score on t^{C} to one of the tables. Hence, t^{C} is the classification variable. Next, in every such table, classify every person to the left column if her score pattern on J^{C} is in $\Omega(t^{\mathrm{C}}, 1)$. Otherwise, assign her to the right column. Next, choose a criterion G; for example, the criterion may be external, as in the classification male/female. But the criterion can also be based on the response patterns from the item set J^{R}; for example, if only one item i is in J^{R}, one may choose $G = 1$ if $X_i = 1$ and $G = 0$ otherwise.

This illustrates the R/C notation: The item set J^{C} is used to construct the columns of the MH table, and participants with a response pattern $\mathbf{w}^{\mathrm{C}} \in \Omega(t^{\mathrm{C}}, 1)$ are assigned to the first column. Hence, the set J^{C} is used to perform the column classification. Also, the variable t^{C} is used to segregate the MH tables, that is, for the MH table classification. Furthermore, the set J^{R} can (but does not have to) be used to construct the row headings, and can thus serve to perform the row classification.

Now we may state the following: under the Rasch model,

$$\Pr(\mathbf{W}^{\mathrm{C}} \in \Omega(t^{\mathrm{C}}, 1) | t^{\mathrm{C}}, G = 1) = \Pr(\mathbf{W}^{\mathrm{C}} \in \Omega(t^{\mathrm{C}}, 1) | t^{\mathrm{C}}, G = 0). \tag{6}$$

The criterion *G may or may not* be based on the item set J^R (but not on the items comprising J^C). Note, however, that the construction of the rows is always based on information non-overlapping with the information for the column classification. Equation (6) is our central result, and it is proven in the Appendix. It shows that the Rasch model predicts independence for a wide class of MH tables, and hence a χ^2 distribution for the MH statistic. Therefore, an appropriate choice of row and column classifications can serve as a test of specific assumptions of the Rasch model.

One special case is the MH DIF application. Here, choose the criterion $G = 0$ or 1 if person p is male or female respectively (or any other criterion external to the test itself); $J^C = J$ (so J^R is an empty set). Furthermore, pattern $\mathbf{w}^C$ is a member of $\Omega(t^C, 1)$ if and only if $\Sigma_i w_i^C = t^C$ and $w_i = 1$, i.e., if the response pattern results in a subscore t^C and has a 1 at the position of item i. Note that this implies that item i has to be included in calculating the score t, as was done by Holland and Thayer (1988) as well. More applications of this general procedure will be given in the following paragraphs.

A final remark is that one of both columns may sometimes remain empty; for example, if t^C is the perfect score, then the only possible pattern $(1, 1, \ldots, 1)$ (t^C ones) will be assigned to the left- or right-hand column, so one column must remain empty. This is harmless, however, since contributions from this t^C value will become zero in both numerator and denominator. If this occurs too often, a loss of power results. However, the problem will not occur for a reasonable choice of the column classification. We now turn to the applications.

4. Testing for unidimensionality: External criterion

Consider a test consisting of two sets of items, for example, verbal and geometrical analogies. It is suspected that males are relatively better at the geometrical items, while females are relatively better at the verbal items. The Rasch model predicts that gender does not establish a preference for one type of items. To be concrete, we construct the set J^C as consisting of all items in J (so J^R is an empty set). For each score level t, an MH table is constructed (so $t^C = t$). The set $J = J^C$ is now partitioned as $(J^{\text{verb}}, J^{\text{geom}})$, where J^{verb} and J^{geom} contain all verbal and geometrical items, respectively. For the row classification, all males are assigned to Group $G = 0$, while females are assigned to $G = 1$. This example is called an *external criterion* test because the row classification is not based on item responses. For the column classification, the patterns $\mathbf{w}$ in which the score on verbal items is higher than or equal to the score on geometrical items are assigned to the set $\Omega(t, 1)$. More precisely, if we split a response $\mathbf{w}$ into $(\mathbf{w}^{\text{verb}}, \mathbf{w}^{\text{geom}})$, a response pattern $\mathbf{w}$ is assigned to $\Omega(t, 1)$ if and only if

$$\sum_i w_i^{\text{verb}} \geq \sum_i w_i^{\text{geom}}. \tag{7}$$

In other words, if the person has a higher score on the verbal items than on the geometrical items, she is assigned to the left column of the MH table. The number of verbal and geometrical items need not be equal, but a more or less equal number will, of course, increase the power of the test. Second, if the number of items is unequal, it may be more meaningful to take the mean instead of the absolute sum score in (7).

Ties can be handled arbitrarily, and in our application we assign them to the first column (the 'verbal' column). The reasoning in the previous paragraph shows that the Rasch model predicts independence and hence a χ^2 distribution of the MH chi-squared statistic. On the

other hand, if men prefer geometrical items to verbal items (in comparison with women), unidimensionality will be violated and the MH chi-squared statistic will turn out to be too high.

A simulation study for this application will now be presented. Two types of persons are considered (e.g., male and female). The test involved will be an $I = 40$ item test, of which 20 are verbal, 20 geometrical items. (The item naming is, of course, arbitrary in a simulation study, but we pursue the verbal/geometrical naming for concreteness.)

Four factors are varied: number of participants (P); magnitude of Rasch model violation (u); item parameters; and ability distribution. First, all item parameters $\beta_i = -\ln(\epsilon_i)$ for this and the following study are shown in Table 2. From this table one may note that verbal and geometrical items either have identical (distributions of) difficulty levels β (see the left-hand column of the table), or have different (distributions of) difficulty levels (see the right-hand column of the table). In the case of different distributions, the 20 verbal items are always at least as easy as the geometrical items. These cases are extreme, of course, and in most real-life examples the separation will not be that clear. However, this condition is introduced because it gives a conservative power estimate: the more similar the item parameters are, the more powerful the test will be.

Second, abilities $\theta = \ln(\xi)$ for male participants always are sampled from an $N(0, 1)$ distribution. Female abilities are either sampled from an $N(0, 1)$ or from an $N(0.5, 1)$ distribution (see the Table 3 headings $\theta_{\text{fem}} \sim N(0, 1)$ and $\theta_{\text{fem}} \sim N(0.5, 1)$, respectively). Third, the number of participants P varies from 100 to 1000; this factor is given in the row headings of Table 3. Fourth, the Rasch model is violated with magnitude u ($u = 0, 0.1$ or 0.5). For a geometrical item, this implies that item difficulties β_i decrease by an amount u when a male person is solving the item. Similarly, females consider verbal items to be easier, so item difficulties decrease by an amount u when a female person is solving the item. If $u = 0$, data are pure Rasch data and the MH chi-squared statistic should be χ^2 distributed with one degree of freedom.

Table 2. Item parameters for the three applications

		Item parameters β	
Criterion	Item set	Same	Different
External	J^{verb}	$-1, -1, -1, -1, -0.5, -0.5,$ $-0.5, -0.5, 0, 0, 0, 0, 0.5, 0.5,$ $0.5, 0.5, 1, 1, 1, 1$	$-1, -1, -1, -1, -1, -1, -1, -1,$ $-0.5, -0.5, -0.5, -0.5, -0.5, -0.5,$ $-0.5, 0, 0, 0, 0$
	J^{geom}	$-1, -1, -1, -1, -0.5, -0.5,$ $-0.5, -0.5, 0, 0, 0, 0, 0.5, 0.5,$ $0.5, 0.5, 1, 1, 1, 1$	$0, 0, 0, 0, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5,$ $0.5, 0.5, 1, 1, 1, 1, 1, 1, 1, 1$
Internal	J^{verb1}	$-1, -1, -0.5, -0.5, 0, 0,$ $0.5, 0.5, 1, 1$	$-1, -1, -0.5, -0.5, -0.5, -0.5,$ $0, 0, 0, 0$
	J^{verb2}	$-1, -1, -0.5, -0.5, 0, 0,$ $0.5, 0.5, 1, 1$	$-1, -1, -0.5, -0.5, -0.5, -0.5,$ $0, 0, 0, 0$
	J^{geom1}	$-1, -1, -0.5, -0.5, 0, 0,$ $0.5, 0.5, 1, 1$	$0, 0, 0, 0, 0.5, 0.5, 0.5, 0.5,$ $1, 1$
	J^{geom2}	$-1, -1, -0.5, -0.5, 0, 0,$ $0.5, 0.5, 1, 1$	$0, 0, 0, 0, 0.5, 0.5, 0.5, 0.5,$ $1, 1$

Table 3. Simulation results for the external criterion MH test

	Item difficulties					
	Same			Different		
	Violation (u)			Violation (u)		
	0	0.1	0.5	0	0.1	0.5
$\theta_{\text{fem}} \sim N(0, 1)$						
$P = 100$	0.043	0.157	0.993	0.030	0.038	0.478
$P = 500$	0.044	0.648	*	0.044	0.200	*
$P = 1000$	0.048	0.930	*	0.049	0.369	*
$\theta_{\text{fem}} \sim N(0.5, 1)$						
$P = 100$	0.054	0.142	0.994	0.025	0.039	0.500
$P = 500$	0.052	0.659	*	0.051	0.211	*
$P = 1000$	0.046	0.924	*	0.043	0.366	*

For each factor combination, 1000 data sets are generated and the MH chi-squared statistic is calculated for each data set. We report the proportion of data sets that is rejected at the $\alpha = 0.05$ level. An asterisk (*) is written if all 1000 data sets are rejected. Theoretical expectations are as follows. All cells with $u = 0$ should have a rejection proportion of about 0.05. Notice that this holds independent of ability distributions or item difficulties, as was stated above and proven in the Appendix. However, these factors may be expected to influence the power of the test: unequal θ distributions result in a loss of power since males and females will less often be found in the same *MH* table (i.e., will less often have the same value on the classification variable t). Similarly, unequal item difficulties result in a tendency for data to be concentrated in one column of the *MH* table, which again lowers power. That the number of participants (P) and size of violation (u) factors influence the power of the test is obvious.

Most of these predictions are shown to be valid in Table 3. Notice, however, that unequal ability distributions hardly (if at all) influence the power of the test statistic. Furthermore, one may note that even very small deviations ($u = 0.1$) can be detected by this test if the sample size (P) is sufficiently large. On the other hand, with small $P(P = 100)$, the test has only reasonable power in the condition with a reasonably strong violation and equal item difficulties. In the following section, the same example will be pursued.

5. Testing for unidimensionality: Internal criterion

Let us now assume that no good *a priori* classification is known; nevertheless, it is suspected that some people are better at one part of the test, some better at another (again, assume the verbal/geometrical distinction can be made). To test for this, again make the division ($\mathbf{w}^{\text{verb}}, \mathbf{w}^{\text{geom}}$), but then split each part again in two, so that the complete response pattern is now $\mathbf{w} = (\mathbf{w}^{\text{verb1}}, \mathbf{w}^{\text{verb2}}, \mathbf{w}^{\text{geom1}}, \mathbf{w}^{\text{geom2}})$. The set J^{R} consists of all items involved in $\mathbf{w}^{\text{verb1}}$ and $\mathbf{w}^{\text{geom1}}$. The set J^{C} contains all items involved in the response patterns $\mathbf{w}^{\text{verb2}}$ and $\mathbf{w}^{\text{geom2}}$. This test is called an *internal criterion* test because the row classification is based on item

responses. All persons with the same score on the variable

$$t^{\mathrm{C}} = \sum_i w_i^{\mathrm{verb2}} + \sum_i w_i^{\mathrm{geom2}}$$

are assigned to the same MH table (for score t^{C}). In such a table, a person with score t^{C} is assigned to row 1 (i.e., $G = 1$) if

$$\sum_i w_i^{\mathrm{verb1}} \geq \sum_i w_i^{\mathrm{geom1}}$$

and to row 2 ($G = 0$) otherwise. A person is assigned to column 1 if

$$\sum_i w_i^{\mathrm{verb2}} \geq \sum_i w_i^{\mathrm{geom2}}$$

and to column 2 otherwise. Note that, as always, rows and columns are constructed based on non-overlapping information.

As before, it is not necessary that all parts contain the same number of items, but the same arguments about loss of power and about taking means instead of sums apply here. Under the Rasch model, there should be no association between the row heading 'better at verbal material in the first part' and the column heading 'better at verbal material in the second part' (or vice versa, of course). On the other hand, if some people are relatively better at verbal items, some better at geometrical, this will be detected by the procedure constructed above. In contrast with the previous example, it is not known *a priori* which person belongs to which group. However, *a priori* knowledge about the items is still required.

Some simulation results for this test will now be described. Again, the test is an $I = 40$ item test and $P = 100, 500$ or 1000 (see Table 4). Two types of persons are considered, denoted the verbal group and the geometrical group, respectively. Abilities of the geometrical group are $N(0, 1)$ distributed, while abilities in the verbal group are either $N(0, 1)$ or $N(0.5, 1)$ distributed (see the column headings in Table 4; abilities of the verbal group are here denoted by θ_{verb}). Rasch model violation (u) is defined as follows. If a person (with constant ability θ_p) belonging to the verbal group is solving a verbal item, the difficulty of this item decreases to a value $\beta_i - u$ (of course, it again increases to β_i when solving a geometrical item). Similarly, every parameter β_i of a geometrical item decreases to a value $\beta_i - u$ when a person from the geometrical group is solving the item. Again, the case $u = 0$ denotes the case in which the Rasch model is not violated. Denote the verbal items in J^{R} by J^{verb1}, the verbal items in J^{C} by J^{verb2}, and similarly for geometrical items. The corresponding item parameters can be found in Table 2. Again, there is the case in which verbal and geometrical items are equally distributed, and the case in which geometrical items are never easier. As in the previous simulation study, the latter case is expected to be less powerful.

Results for this application are shown in Table 4. A major difference from the previous example is that the power is lower in general. Presumably, this is because the column classification is now based on 20 items only (instead of 40). However, for model violations that are sufficiently large and for a not too small P the MH test detects that the item set is not unidimensional.

A test that can be used to investigate the same hypothesis is the Martin-Löf (ML) test described above. Hence, it seems useful to compare the present simulation results with results from the ML test. We resimulate data from cell ($\theta_{\mathrm{verb}} \sim N(0, 1), u = 0, P = 500$) of Table 4,

Table 4. Simulation results for the internal criterion MH test

	Item difficulties					
	Same			Different		
	Violation (u)			Violation (u)		
	0	0.1	0.5	0	0.1	0.5
$\theta_{\text{verb}} \sim N(0, 1)$						
$P = 100$	0.051	0.060	0.253	0.034	0.042	0.131
$P = 500$	0.047	0.059	0.890	0.054	0.043	0.512
$P = 1000$	0.055	0.051	0.993	0.062	0.044	0.793
$\theta_{\text{verb}} \sim N(0.5, 1)$						
$P = 100$	0.057	0.035	0.225	0.038	0.035	0.128
$P = 500$	0.048	0.065	0.841	0.054	0.061	0.487
$P = 1000$	0.059	0.050	0.991	0.046	0.054	0.756

where item difficulties for both items sets are the same. The ML procedure requires that each item set is split into two subsets and estimation is done in each subset separately. We chose the verbal/geometrical distinction to split the items as this results in a test of unidimensionality (see Gustafsson, 1980). The mean ML value was equal to 235.185, while the mean of the corresponding chi-square distribution is equal to 399 ($= 20^2 - 1$); clearly the chi-square approximation is not good. Zero data sets (out of 1000) were rejected, while the expected number is 50 at level $\alpha = 0.05$. Similarly, if we change the parameter u to 0.5 (and all other parameters are unchanged), zero data sets were rejected, while the corresponding MH test has large power (890 out of 1000 data sets rejected; see Table 4).

We lower the number of items in a data set in order to make the chi-squared approximation

Table 5. Comparison of (internal criterion) MH and ML test

	Procedure					
	Mantel–Haenszel			Martin-Löf		
	Violation (u)			Violation (u)		
	0	0.5	1	0	0.5	1
$\theta_{\text{verb}} \sim N(0, 1)$						
$P = 100$	0.049	0.081	0.357	0.028	0.050	0.330
$P = 500$	0.063	0.178	0.960	0.043	0.128	0.993
$P = 1000$	0.047	0.321	*	0.050	0.226	*
$\theta_{\text{verb}} \sim N(0.5, 1)$						
$P = 100$	0.048	0.074	0.362	0.018	0.036	0.331
$P = 500$	0.047	0.176	0.938	0.032	0.136	0.996
$P = 1000$	0.042	0.291	0.998	0.051	0.426	*

more accurate. There are now six verbal and six geometrical items (so $I = 12$ instead of $I = 40$), with parameter vector $\boldsymbol{\beta} = (-1, 0, 1, -1, 0, 1)$ for both sets. This lower number of items may result in a lower power. Hence, in order to see the full range of the 'power spectrum' we enlarged the violation size u to $u = 0$, 0.5 or 1. The results comparing the two procedures (MH and ML) on the same data sets are shown in Table 5. One can see that both tests have similar properties. Both are chi-squared distributed under the Rasch model; when the model is violated, sometimes the power of one test is higher, sometimes the other, possibly reflecting just random fluctuations. Hence, the conclusion of the comparison would be that, if the number of items is relatively large, the MH test presented here does better. Otherwise, the ML test is appropriate also.

6. Item dependence

Under local stochastic independence, any two items are independent given the latent trait θ: the score on an item does not contain a clue to the score on other items. In reality, it may well be that two items are dependent given θ. This would occur, for example, if solving one item is dependent on the answer obtained in the previous item. Conditional association in this sense may also be seen as a violation of unidimensionality. For example, Van den Wollenberg (1982) developed a procedure based on this idea to test for unidimensionality (his Q_2 statistic). We present a test to investigate two-item dependence which is a member of the class discussed above.

Suppose two items j and k are suspected to be associated conditional on θ. The item set is partitioned into $J^{\mathrm{R}} = \{\text{item } j\}$, and $J^{\mathrm{C}} = \{\text{item } 1, \ldots, \text{item } j-1, \text{item } j+1, \ldots, \text{item } I\}$; the corresponding partitioning within response patterns is $(w_j, \mathbf{w}_{-j})$ where $\mathbf{w}_{-j}$ denotes the response pattern $\mathbf{w}$ in which item j is deleted. For example, if $j = 1$, then $\mathbf{w}_{-1} = (w_2, \ldots w_I)$. Further, define

$$t^{\mathrm{C}} = \sum_{i \neq j} w_i. \tag{8}$$

The MH classification variable is t^{C} as defined in (8), that is, the sum score where item j is deleted. This follows from the fact that G already 'uses' the item j, so t^{C} can no longer include this item as well. The column classification is based on item k: every person solving k correctly

Table 6. Simulation results for the item dependence MH test

	Item difficulties					
	Same			Different		
	Violation (u)			Violation (u)		
	0	0.1	0.5	0	0.1	0.5
$\theta \sim N(0,1)$						
$P = 100$	0.053	0.053	0.128	0.063	0.058	0.139
$P = 500$	0.047	0.084	0.614	0.053	0.067	0.582
$P = 1000$	0.051	0.100	0.925	0.028	0.079	0.887

is assigned to the left column, all others to the right one. The row classification is based on item j: persons solving j correctly are assigned to the upper row, other persons to the lower one. As shown above, such a table is not associated under the Rasch model. On the other hand, if there is some covariance left between these items, the MH chi-squared statistic will be too large.

It is useful to make a link here with the procedure of Rosenbaum (1984) discussed above. If one were to apply Rosenbaum's procedure, the MH classification variable would consist of the sum score based on all items except items j and k; the row and column classifications would look exactly the same. Correlation larger than or equal to zero would be predicted under the model (and tested for). A difference from our approach is that Rosenbaum's procedure was derived assuming two general assumptions, monotonicity and conditional independence, rather than a specific model.

A simulation study is now set up for the item dependence MH test. Sample and test sizes are as before (i.e., $I = 40$ and $P = 100$, 500 or 1000). Item parameters β can be found in the upper left cell of Table 2 (see the headings 'Criterion–external' and 'Item parameters–Same'). Abilities θ are sampled from a standard normal distribution. We either compare two items j and k with $\beta = 0$ (under the heading 'Item difficulties–Same' in Table 6) or two items j and k with β_j and β_k equal to -0.5 and 0.5, respectively (under the heading 'Item difficulties–Different' in Table 6). Violation of the Rasch model is implemented as follows: people with $X_j = 1$ solve item k with item difficulty $\beta_k - u$ instead of the usual β_k (note that this implies $j < k$). The results are displayed in Table 6. The power seems less than for the two previous examples. The test now is only powerful if both violation size and sample size (P) are reasonably large. One plausible reason is that violation is only unidirectionally defined in this example, in the sense that solving item j is beneficial for solving item k, but not solving item j (i.e., belonging to $G = 0$) does not influence the ability. Thus, model violations are not detected when the violation is only small.

7. Illustration

As an illustration of this method, we analyse a data set of 441 participants, who have completed a test consisting of 78 questions on diverse topics. The test was an assessment test (we will denote it the Law Entry Test) for aspiring law students: the subjects covered were courses they would take in their curriculum, such as law, sociology and psychology. In total, there were 11 topics, with six to ten questions on each topic. The global R_{1c} Rasch model test (Glas, 1988) was applied to these data, which turned out to be significant ($p = 0.000$). This test, however, does not provide much specific information about which model assumptions are violated. On the other hand, our MH testing approach is capable of testing many specific aspects of the model, as will now be illustrated.

One obvious way in which unidimensionality may be violated is that different courses assess different abilities. This may be evaluated by comparing items from different courses. Suppose we wish to compare the psychology and the law parts of the Law Entry Test. In this fictional example, items 1–6 are law items and 7–12 psychology items. In each such table are grouped those participants who have the same score on the variable

$$X_4 + X_5 + X_6 + X_{10} + X_{11} + X_{12},$$

to which may be added the score of additional items if desired (items 13, 14, ...), but not items 1, 2, 3, 7, 8 and 9, because they are used in the row classification. This row classification

is based on the score on the first part of the psychology and law items, the column classification on the second part. More specifically, we assign to row 1 everyone with

$$x_1 + x_2 + x_3 \geq x_7 + x_8 + x_9,$$

while others are assigned to row 2. Similarly, we assign to column 1 everyone with

$$x_4 + x_5 + x_6 \geq x_{10} + x_{11} + x_{12},$$

while others are placed in column 2. An important consideration here is that to be maximally powerful items (4, 5, 6) should be of comparable difficulty to items (10, 11, 12), and items (1, 2, 3) should be comparable to (7, 8, 9). Moreover, if the number of items on both sides of the equation is not equal, it is more sensible to take means instead of sums, as was noted earlier. For example, if the items 4 and 5 are assigned to column 1, and the items 10, 11 and 12 to column 2, it is allowed to classify response patterns according to

$$x_4 + x_5 \geq x_{10} + x_{11} + x_{12},$$

but since a loss of power is likely to be the result, it is more sensible to classify response patterns according to

$$\frac{x_4 + x_5}{2} \geq \frac{x_{10} + x_{11} + x_{12}}{3},$$

which of course reduces to the previous expression if the numbers of variables on each side of the equation are equal. This approach (i.e., taking means instead of sums) will also be pursued in the following.

We will compare the law and psychology parts of the test using this procedure. These parts consist of six and eight items, respectively. The first three items of the law test and the first four items of the psychology test are used to construct the row classification as described above. That is, if the mean score on the law subtest is higher than the mean score on the psychology subtest, a person is assigned to the first row; otherwise, she is assigned to the second row. A similar procedure is followed for the remaining items in the column classification. This results in an MH value of 10.820 ($df = 1; p < 0.01$). Similarly, the Martin-Löf test rejects the model (ML = 175.97; $df = 47; p < 0.01$). The reason for this multidimensionality might be that some very specific ability is shared by only a few items within one of the parts, such as knowledge of a specific fact, resulting in item dependence. To test whether this phenomenon occurs, we have to construct the MH table in a different way. In any such table we aggregate those participants who have achieved a score of t, in which item j is excluded from the calculation. We take the law subtest and compare some of the items

Table 7. Some comparisons between the Law Entry Subtest questions

Question numbers	MH	p-value
73–76	1.159	0.28
73–77	1.264	0.26
73–78	22.778	<0.01
76–77	3.185	0.07
76–78	0.956	0.32
77–78	6.290	0.01

because previous research has indicated that this test may contain item dependencies (Tuerlinckx & De Boeck, 1998). The results are displayed in Table 7. As can be seen some dependencies exist between the items which cannot be explained by the model. In particular, the relation between items 73 and 78 seems problematic. Upon inspection of these two items, one can see that they both require knowledge of a specific fact, namely that the government in Belgium is stronger than the parliament.

Although construction of a new model is beyond the scope of this paper, we may perform a new analysis in which item 78 is removed. This item is believed to be worse than item 73 because it is also rather strongly correlated with item 77 (see Table 7), while item 73 is not. If we again apply the MH unidimensionality test to the law and psychology subtests, we end up with a MH value of 3.559 ($p = 0.06$), so now the hypothesis of unidimensionality is no longer rejected.

8. Discussion and conclusion

In this paper, we have constructed a general class of Rasch model tests, from which various examples may be derived according to the wishes of the investigator. Standard (or easily implemented) MH software can then be used to test different Rasch model assumptions. However, the simulation results described above indicate that, in some applications, violations will only be detected if the violation and the sample size are large enough. The Law Entry Test illustration shows that the method is useful for detecting and testing the occurrence of specific violations.

The set of tests we have proposed can be characterized on several dimensions. First, model tests may or may not use estimated (item or person) parameters in the calculation. This aspect distinguishes parametric from non-parametric tests. The present class of tests is non-parametric in this sense; this makes it (relatively) easy to apply the test and derive results concerning its distribution.

Secondly, tests may or may not have a known (asymptotic) distribution under the null hypothesis. The tests we have proposed all possess this property. The same holds, for example, for Glas's (1998) R_{1c} test and Rosenbaum's set of test statistics.

Thirdly, we have constructed a whole set of tests all based on the same theorem (i.e., equation (6)). The Holland and Thayer (1988) item bias test is a special case of this class. Other general classes of Rasch model tests were proposed by Glas (1988) and Rosenbaum (1984). Such a general class is useful since it allows the testing of some hypotheses which were difficult to test directly before. For example, the external criterion/unidimensionality test described above seems to be a test for an interesting hypothesis which was difficult to evaluate previously.

Acknowledgements

We wish to thank Eric Maris, Patrick Onghena and Gert Storms for their useful comments and Laurence Claes and Piet J. Janssen for the use of their data.

References

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

Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests.* [Introduction to psychological test theory.] Berne: Huber.
Fischer, G. H. (1993). Notes on the Mantel–Haenszel procedure and another chi-squared test for the assessment of DIF. *Methodika, 7,* 88–100.
Fischer, G. H. (1995a). Derivations of the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 15–38). New York: Springer-Verlag.
Fischer, G. H. (1995b). Some neglected problems in IRT. *Psychometrika, 60,* 459–487.
Glas, C. A. W. (1988). The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika, 53,* 525–546.
Glas, C. A. W., & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 69–95). New York: Springer-Verlag.
Gustafsson, J. (1980). Testing and obtaining fit of data to the Rasch model. *British Journal of Mathematical and Statistical Psychology, 33,* 205–233.
Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel–Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity,* Hillsdale, NJ: Erlbaum.
Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22,* 719–748.
Mood, A. M., Graybill, F. A., & Boes, D. C. (1974). *Introduction to the theory of statistics,* Singapore: McGraw-Hill.
Parshall, C. G., & Miller, T. R. (1995). Exact versus asymptotic Mantel–Haenszel DIF statistics: A comparison of performance under small-sample conditions. *Journal of Educational Measurement, 32,* 302–316.
Ponocny, I. (1999). *Non-parametric goodness-of-fit tests for the Rasch model.* Unpublished manuscript.
Ponocny, I., & Ponocny-Seliger, E. (1999). *T-Rasch 1.0.* Groningen: ProGAMMA.
Rasch, G. (1966). An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology, 19,* 49–57.
Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika, 49,* 425–436.
Tuerlinckx, F., & De Boeck, P. (1998). *The effect of ignoring local item dependencies on the estimated discrimination parameters.* Research Report 98-2. University of Leuven, Research Group on Quantitative Methods and Human Intelligence.
Uttaro, T., & Millsap, R. E. (1994). Factors influencing the Mantel–Haenszel procedure in the detection of differential item functioning. *Applied Psychological Measurement, 18,* 15–25.
Van den Wollenberg, A. L. (1982). Two new test statistics for the Rasch model. *Psychometrika, 47,* 123–140.
Verhelst, N. D., Glas, C. A. W., & Van der Sluis, A. (1984). Estimation problems in the Rasch model: The basic symmetric functions. *Computational Statistics Quarterly, 1,* 245–262.
Zwick, R. (1990). When do item response function and Mantel–Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics, 15,* 185–197.
Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30,* 233–251.

Received 22 April 1999; revised version received 11 February 2000

Appendix

Consider the left-hand side of equation (6):

$$\Pr(\mathbf{W}^{\mathrm{C}} \in \Omega(t^{\mathrm{C}}, 1) | t^{\mathrm{C}}, G = 1) = \sum_{\mathbf{w}^{\mathrm{C}} \in \Omega(t^{\mathrm{C}},1)} \Pr(\mathbf{w}^{\mathrm{C}} | t^{\mathrm{C}}, G = 1) = \sum_{\mathbf{w}^{\mathrm{C}} \in \Omega(t^{\mathrm{C}},1)} \frac{\Pr(\mathbf{w}^{\mathrm{C}} | G = 1)}{\Pr(t^{\mathrm{C}} | G = 1)},$$

since $\Pr(\mathbf{w}^{\mathrm{C}}, t^{\mathrm{C}}|G=1) = \Pr(\mathbf{w}^{\mathrm{C}}|G=1)$: The score t^{C} is simply the sum of the item scores in $\mathbf{w}^{\mathrm{C}}$. This can be elaborated as

$$\sum_{\mathbf{w}^{\mathrm{C}} \in \Omega(t^{\mathrm{C}},1)} \frac{\int_{\xi} \Pr(\mathbf{w}^{\mathrm{C}}|\xi, G=1)\mathrm{d}F(\xi|G=1)}{\int_{\xi} \Pr(t^{\mathrm{C}},|\xi, G=1)\mathrm{d}F(\xi|G=1)}. \tag{A1}$$

Now consider an arbitrary numerator of a term in equation (A1),

$$\int_{\xi} \Pr(\mathbf{w}^{\mathrm{C}}|\xi, G=1)\mathrm{d}F(\xi|G=1),$$

which can be expanded as

$$\int \prod_{i \in J^{\mathrm{C}}} \frac{(\xi\epsilon_i)^{w_i^{\mathrm{C}}}}{1+\xi\epsilon_i}\mathrm{d}F(\xi|G=1), \tag{A2}$$

since the group membership G is not based on the items in J^{C} and since we have conditioned on ξ (from which the local stochastic property can be applied). Furthermore, in (A2), w_i^{C} is the response on item i in response pattern $\mathbf{w}^{\mathrm{C}}$. Bringing the ξ-independent part outside of the integral sign, we can write (A2) as

$$\prod_{i \in J^{\mathrm{C}}} \epsilon_i^{w_i^{\mathrm{C}}} \int \xi^{t^{\mathrm{C}}} \prod_{i \in J^{\mathrm{C}}} \frac{1}{1+\xi\epsilon_i}\mathrm{d}F(\xi|G=1). \tag{A3}$$

On the other hand, a denominator of a term in (A1) can be written as

$$\sum_{\mathbf{w}^{\mathrm{C}} \in \Omega(t^{\mathrm{C}})} \int \Pr(\mathbf{w}^{\mathrm{C}}|\xi, G=1)\mathrm{d}F(\xi|G=1), \tag{A4}$$

where the summation is taken over all response patterns resulting in the sum score t^{C} (which was earlier defined as $\Omega(t^{\mathrm{C}})$). Hence, (A4) can be rewritten, along the lines of the derivation just given, as

$$\begin{aligned}&\sum_{\mathbf{w}^{\mathrm{C}} \in \Omega(t^{\mathrm{C}})} \prod_{i \in J^{\mathrm{C}}} \epsilon_i^{w_i^{\mathrm{C}}} \int \xi^{t^{\mathrm{C}}} \prod_{i \in J^{\mathrm{C}}} \frac{1}{1+\xi\epsilon_i}\mathrm{d}F(\xi|G=1) \\ &= \int \xi^{t^{\mathrm{C}}} \prod_{i \in J^{\mathrm{C}}} \frac{1}{1+\xi\epsilon_i}\mathrm{d}F(\xi|G=1) \sum_{\mathbf{w}^{\mathrm{C}} \in \Omega(t^{\mathrm{C}})} \prod_{i \in J^{\mathrm{C}}} \epsilon_i^{w_i^{\mathrm{C}}}.\end{aligned} \tag{A5}$$

Since the integrals in (A3) and (A5) are of the same form, dividing (A3) by (A5) gives

$$\frac{\prod_{i \in J^{\mathrm{C}}} \epsilon_i^{w_i^{\mathrm{C}}}}{\gamma_{t^{\mathrm{C}}}(\epsilon)},$$

where $\gamma_t(.)$ denotes the elementary symmetric function of order t (Verhelst, Glas & Van der Sluis, 1984). The formula (A1) is then equal to

$$\sum_{\mathbf{w}^{\mathrm{C}} \in \Omega(t^{\mathrm{C}},1)} \frac{\prod_{i \in J^{\mathrm{C}}} \epsilon_i w_i^{\mathrm{C}}}{\gamma_{t^{\mathrm{C}}}(\epsilon)},$$

and so $\Pr(\mathbf{W}^{\mathrm{C}} \in \Omega(t^{\mathrm{C}}, 1) | t^{\mathrm{C}}, G = 1)$ is independent of the variable G (which only depends on external variables or the item set J^{R}) and the result follows.