Psicologica (2008), 29, 229-254

Some Considerations on the Partial Credit Model

N.D. Verhelst*and H.H.F.M. Verstralen
National Institute for Educational Measurement (Cito)
Arnhem, The Netherlands

Abstract

The Partial Credit Model (PCM) is sometimes interpreted as a model for stepwise solution of polytomously scored items, where the item parameters are interpreted as difficulties of the steps. It is argued that this interpretation is not justified. A model for stepwise solution is discussed. It is shown that the PCM is suited to model sums of binary responses which are not supposed to be stochastically independent. As a practical result, a statistical test of stochastic independence in the Rasch model is derived.

1 Introduction

Masters (1982) introduced the *partial credit model (PCM)* as an IRT model for polytomous items with ordered categories. The rationale he used to introduce the model was based on a response process where the subject responds sequentially to a number of subproblems in the item. The partial credit given equals the number of steps completed successfully, which of course in this rationale should be the first steps. This rationale, together with the tempting conclusion that the location

*Correspondence to: N. D. Verhelst, Cito, P.O. Box 1034, NL-6801 MG, Arnhem, The Netherlands

parameters in the PCM could be interpreted as difficulty parameters of the respective steps, was criticized by Molenaar (1983), who argued that the steps interpretation in the PCM is not justified.

This leaves two important questions:

1. If the PCM is not suited as a formalization of the steps rationale, does there exist other models which can be used for this purpose?

2. Does there exist a compelling rationale that justifies the use of the PCM?

The first question will be addressed briefly in Section 2, where it is explained in some detail why the steps interpretation is not justified in the PCM and where another model, especially designed to allow for such an interpretation is discussed.

The second question, however, is the central focus of the present article: it investigates the relation between the Rasch model and the PCM. This is done in a number of stages. In the first stage (Section 3) it is shown that if a test complies to the Rasch model it also complies to the PCM in the sense that subsets of the items, called testlets, are considered as polytomous items with a score equal to the sum score on the items in the testlet. The converse, however, does not hold: if response patterns consisting of testlet scores comply to the PCM, it does not follow that the Rasch model holds at the level of the individual items, or more generally: the PCM is a much more general model than the Rasch model.

In the next stage (Section 4), a general model for binary items is introduced, where it is possible to allow for a large number of interactions. The Rasch model is a special case of this general family. In the Rasch model all interactions vanish, and consequently it is the unique member of this family where conditional independence between all item responses exist. Two theoretical results are presented for the relation between this model and the partial credit model, applied to testlet scores. The first result (Section 4.1) is that each member of this family complies to the

PCM and the second result (Section 4.2) says that every PCM applied to testlet scores can be considered as a model for sums of binary item scores and thus complies to the general dependence model. The scientific relevance of this finding resides in the fact that the PCM is suitable model for tests of binary items where the condition of local independence is not met, without the necessity to explicitly model the precise form of the interaction effects.

In Section 5, two practical implications of this approach are investigated. The first gives an answer to the question whether in estimating individual abilities of test takers, information is lost if the partial credit model is used in case the Rasch model holds (Section 5.1). The second implication relates to a general condition that has to be fulfilled for the results of Section 4 to be valid. This condition is that testlet scores must be locally independent. In Section 5.2 two methods are discussed to create testlets where there is within testlet dependency but no between testlet dependency.

The article is concluded by a discussion section.

2 The step interpretation of the Partial Credit model

The definition of the PCM states that for an item with maximum score m,

$$P(X = j|\theta, X = j \text{ or } X = j-1) = \frac{\exp(\theta + \beta_j)}{1 + \exp(\theta + \beta_j)}, \tag{1}$$

where θ is the latent variable, and X the item score with values $j = 0, \ldots, m$. The parameters β_j denote the m parameters associated with the item. Now suppose we construct the following two-step item

$$\text{two step item: } \frac{1/2 + 0.25}{0.03} = ?$$

which of course will lead to a completely correct response only if the first step (the addition) and the second step (the division) are computed correctly. We can embed this item into a three-step item, where the third step can only be applied if the first two steps are completed. We can also vary the difficulty of the third step, which we do as an example in the following three versions of the three step item:

version A: $\frac{1/2+0.25}{0.03} + 1 = ?$

version B: $\sqrt{\frac{1/2+0.25}{0.03}} = ?$

version C: $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} (x - \frac{1/2+0.25}{0.03}) \exp(-x^2/2) dx = ?$

For 15 year old students, we may safely say that step 3 in version A is trivially simple, while the third step of version C will be extremely difficult, and will be solved only by a few mathematically gifted students. The third step of Version B is probably not trivially easy in that age group, but one can assume that a substantial proportion of the population masters the concept of the square root function. The step interpretation of the PCM implies that the value of β_1 and β_2 will be equal for the three versions of the three step item. But this is not consistent with (1) as will be shown by the following example, where we concentrate on β_2 and on the item versions B and C.

Consider the population of all persons with $\theta = \theta_0$. In view of the interpretation given to the items, the response probabilities in Table 1 might hold. Note that the probabilities of obtaining a score of 0, 1 and (2 or more) are the same for both versions; in version B, however, 2/3

Table 1: Response probabilities at $\theta = \theta_0$

score:	0	1	2	3
version B	0.1	0.45	0.15	0.3
version C	0.1	0.45	0.44999	0.00001

of the students having reached successfully step 2, can also solve step 3, while in version C almost nobody is successful on step 3. For version C, the probability of a score of 2, given that the score is 1 or 2 is very close to one half, whence it follows from (1) that β_2 will be very close to $-\theta_0$. In version B, however, the conditional probability of obtaining a score of 2, given that the score is 1 or 2 is 0.25, whence it follows, using (1), that $\beta_2 = -(\theta_0 + \ln 3)$. This shows that the value of β_2 does not depend uniquely on the difficulty of the second step but also on the difficulty of the subsequent step(s), and consequently that any interpretation of PCM parameters as difficulties of specific item steps is void.

The conclusion is that the PCM is not suitable to model sequential solution strategies. An appropriate model was found independently at two different places at about the same time. De Vries (1988) and Verhelst, Glas and De Vries (1997) developed a model by combining the simple Rasch model with a subject controlled incomplete design: the steps or subitems of a polytomously scored item are conceived as being administered in a fixed sequence and the next subitem is presented if and only if the previous one is correctly responded to. The answer to each subitem is modeled by the simple Rasch model. The presentation of a subitem thus depends on the behavior of the responding subject, hence the qualification subject controlled. Tutz (1990, 1997) followed the same rationale, but introduced the model formally and more generally as

$$p_j \equiv P(X > j|\theta, X \geq j) = F(\theta + \beta_j), \quad (j = 0, \ldots, m-1), \tag{2}$$

where $F(.)$ is an arbitrary distribution function. It can readily be seen that in both models, the category response functions are given by

$$P(X = j|\theta) = \begin{cases} (1 - p_j) \prod_{g=0}^{j-1} p_g & \text{if } j < m, \\ \\ \prod_{g=0}^{m-1} p_g & \text{if } j = m, \end{cases} \tag{3}$$

whence it follows that both models are identical if F is the logistic distribution function with argument $\theta + \beta_j$

3 The distribution of sums of Rasch item scores

Suppose $m(>1)$ items can be described by the Rasch model, i.e. for any value of the latent variable θ,

$$P(Y_i = y_i|\theta) \propto \exp[y_i(\theta + \beta_i)], \quad (i = 1, \dots, m), \tag{4}$$

where $y_i \in \{0, 1\}$.

Defining the variable S as $S = \sum_i Y_i$, and assuming conditional independence as usual, it is readily seen that

$$P(S = s|\theta) \propto \exp(s\theta) \sum_{\Sigma y = s} \prod_i \varepsilon_i^{y_i}, \tag{5}$$

where $\varepsilon_i = \exp(\beta_i)$. The combinatorial function represented by the sum in the right-hand side of (5) is known as the basic or elementary symmetric function (of order s) of the multivariate argument $\varepsilon = (\varepsilon_1, \dots, \varepsilon_m)$, and will be denoted by $\gamma_s(\varepsilon)$. It is defined formally as

$$\gamma_s \equiv \gamma_s(\varepsilon) = \sum_{\Sigma y = s} \prod_i \varepsilon_i^{y_i}, \quad (s = 0, \dots, m). \tag{6}$$

Note that $\gamma_0(\varepsilon) = 1$. Defining

$$\eta_s = -\ln \gamma_s(\varepsilon), \quad (s = 0, \dots, m), \tag{7}$$

equation (5) can be rewritten as

$$P(S = s|\theta) \propto \exp(s\theta - \eta_s), \tag{8}$$

which is nothing more than the category response function of the PCM in a parameterization first used by Andersen (1977). Notice that η_0 equals zero.

Suppose that a test that consists of k Rasch items is partitioned into T classes, consisting of $m_1, \dots, m_T$ items. These classes will be called *testlets*, and the sums of the item scores in each testlet will be called *testlet scores*. The distributions of these testlet scores can be described by the PCM because the original item responses are independent and the classes are disjoint.

There are two important observations to be made in connection with this result. First, if only testlet scores are observed instead of the original item scores, then it is in principle possible - although not easy - to estimate the original Rasch parameters from the sum scores. The problem to be solved in case of Maximum Likelihood (ML) estimation is this: find the values of the PCM parameters η that maximize the likelihood under the restriction that for each testlet t there exist m_t positive *real* numbers $\varepsilon_{t1}, \ldots \varepsilon_{tm_t}$ such that the non-linear restrictions given by (7) hold for each testlet. It these η-values are found, the ε-parameter estimates may be found from solving for each testlet the system of non-linear equations given by (7). But, even when one succeeds in finding ML-estimates for the ε-parameters, it is not possible to associate them with the original items. If all m_t ε-parameters are distinct in testlet t, then there are $m_t!$ different associations possible, and there is no way of deciding between them on the basis of the testlet scores alone.

The second observation is more important. Although it is true that sums of Rasch item scores are distributed acording to the PCM, the converse is not true: polytomous item scores whose distribution is given by the PCM can*not* always be interpreted as sums of Rasch item scores. If they were, it would follow that for m arbitrary numbers $\eta_1, \ldots, \eta_m$, there would exist m (positive) real numbers $\varepsilon_1, \ldots, \varepsilon_m$ such that (7) is true, and this would be equivalent to claiming that all m-th degree polynomials with positive coefficients have m real-valued (negative) roots, which is not true. This is why the ML estimation procedure loosely described in the previous paragraph is difficult. We explain this in more detail.

Consider the polynomial of the m-th degree

$$P_m(x) = \prod_{i=1}^{m}(x + \varepsilon_i), \tag{9}$$

with all ε_i real and positive. Obviously, the roots are real and all negative (they equal $-\varepsilon_i$). Expanding (9) gives

$$P_m(x) = \gamma_0 x^m + \gamma_1 x^{m-1} + \gamma_2 x^{m-2} + \cdots + \gamma_m x^0, \tag{10}$$

where the coefficients $\gamma_s, (s = 0, \ldots, m)$ denote the elementary symmetric functions as defined by (6). Finding the values of ε from the coefficients of the polynomial is equivalent to finding its roots. Determining from the coefficients whether and how many real roots do exist is an unsolved (and probably unsolvable) problem. A necessary condition for the existence of m real roots has been derived by Isaac Newton (Hardy, Littlewood & Pólya, 1952, theorem 51). It is rephrased here as

Theorem 1 *(Newton) If a polynomial P_m as in (10) has real coefficients $\gamma_0, \gamma_1, \ldots, \gamma_m$, then, if there are m real roots, it holds that*

$$\frac{(s+1)(m-s+1)}{s(m-s)}\gamma_{s-1}\gamma_{s+1} \le \gamma_s^2, \quad (s = 1, \ldots, m-1),$$

with equality holding only if all roots are equal.

For $m = 2$, the condition of the theorem is also sufficient for the existence of real roots, but for higher degrees it is not, as the following example shows. Set $\gamma_0, \ldots, \gamma_3$ to $1, 9, 25$ and 17 respectively. It is easily checked that the two inequalities following from the theorem are fulfilled, but the roots of the cubic polynomial are -1, $-4 + i$ and $-4 - i$, i.e., there are two complex roots. Nevertheless, as a necessary condition, the theorem puts severe restrictions on the possibility to interpret PCM item scores as sums of Rasch item scores, since in the PCM no restrictions whatsoever are put on the parameter space; i.e., for a partial credit item with maximum score m, the parameter space is $\mathbb{R}^m$. These restrictions led Van Engelenburg (1997) to the conclusion that the PCM is not an adequate model to describe the distribution of sums of binary item scores. It will be shown in the next section that these restrictions are a direct consequence of assuming local independence between the binary item responses.

4 Models with dependent responses

To model dependencies between item responses, it is easier to model whole response patterns than merely item responses, because dependence means lack of local independence, and therefore impossibility of multiplying item response functions.

As before, we assume that the test consists of k binary items, and is partitioned into T testlets, containing $m_1, \ldots, m_T$ items respectively. As most of the discussion to come will focus on a single testlet, explicit reference to the testlet number will be dropped.

Consider a testlet consisting of $m(> 1)$ items. The vector $\mathbf{Y} = (Y_1, \ldots, Y_m)$ with realizations $\mathbf{y} = (y_1, \ldots, y_m)$ will be called the response pattern. The random variable S, with realizations s, defined by

$$S \equiv S(\mathbf{Y}) = \sum_{i=1}^{m} Y_i, \tag{11}$$

is called the testlet score. Define the m sets I_g, $(g = 1, \ldots, m)$ as the sets containing all ordered g-tuples of the numbers $1, \ldots, m$. This means $I_1 = \{1, \ldots, m\}$, $I_2 = \{(1,2), \ldots, (1,m), (2,3), \ldots, (m-1,m)\}$, etc. The cardinality of I_g is $\binom{m}{g}$. The general model that will be studied is given by

$$P(\mathbf{Y} = \mathbf{y}|\theta) \propto \exp\left[s\theta + \sum_{i\in I_1} y_i\beta_i + \sum_{(i,j)\in I_2} y_iy_j\beta_{ij} + \cdots + \sum_{I_m} y_iy_j\cdots y_m\beta_{ij\cdots m}\right], \tag{12}$$

and by the assumption of local independence between testlet response patterns. Notice that the last sum in the right-hand side of (12) has only one term; it is written as a sum to make the structure of the model clear. The model is a generalization of the Rasch model: if all β-parameters having two or more subscripts are set to zero, the Rasch model results. The extra parameters catch possible interactions between items, and if one of them is non-zero, local independence is lost.

Model (12) and several submodels resulting from setting interaction parameters to zero have been studied by Kelderman (1984); see also Verhelst & Glas (1995a). It should be stressed that this model and various submodels are estimable if the item responses are observed. What matters here, however, is to see what happens if only the testlet scores S_t, $t = 1, \ldots, T$, are observed.

4.1 Testlet scores modeled by the PCM

Since testlet scores are assumed to be independent given θ, it suffices to consider a single testlet (without reference to its number t). Taking the sum of (12) over all response patterns with testlet score s gives

$$P(S = s|\theta) \propto \exp(s\theta)\times \sum_{\Sigma z=s} \exp\left[\sum_{I_1} z_i\beta_i + \sum_{I_2} z_iz_j\beta_{ij} + \cdots + \sum_{I_m} z_iz_j\cdots z_m\beta_{ij\cdots m}\right]. \tag{13}$$

Notice that in the preceding expression the vector $z = (z_1, \cdots, z_m)$ does not refer to any observed response pattern: it is to be understood as the generic expression for a reponse pattern within the testlet. The outer sum in (13) runs over all response patterns having a testlet sum score of s.

To elucidate the structure of Expression 13 and its importance, we write it with another parameterization. Define

$$\varepsilon_i = \exp(\beta_i); \quad \varepsilon_{ij} = \exp(\beta_{ij}); \; \ldots \; ; \; \varepsilon_{ij\cdots m} = \exp(\beta_{ij\cdots m}),$$

and the vector ε^* as

$$\varepsilon^* = (\varepsilon_1, \ldots, \varepsilon_m, \varepsilon_{12}, \ldots, \varepsilon_{m-1,m}, \ldots, \varepsilon_{12\ldots m}).$$

Furthermore, define the combinatorial function $\Gamma_s(\varepsilon^*)$ as

$$\Gamma_s(\varepsilon^*) = \sum_{\Sigma z = s} \prod_{I_i} \varepsilon_i^{z_i} \times \prod_{I_2} \varepsilon_{ij}^{z_i z_j} \times \cdots \times \prod_{I_m} \varepsilon_{ij\cdots m}^{z_i z_j \cdots z_m}, \tag{14}$$

so that (13) can be written as

$$P(S = s|\theta) \propto \exp(s\theta) \times \Gamma_s(\varepsilon^*). \tag{15}$$

For $m = 3$ the sum in the right-hand side of (14) is displayed, term by term, for the three possible patterns that have a score of 1 or 2 (see Table 2). For a score of zero, the sum has one term equal to 1, and for a score of 3, the sum also consists of a single term equal to the product of all ε-parameters.

Table 2: Illustration of (14)

score = 2				score = 1			
pattern			term	pattern			term
1	1	0	$\varepsilon_1\varepsilon_2\varepsilon_{12}$	1	0	0	ε_1
1	0	1	$\varepsilon_1\varepsilon_3\varepsilon_{13}$	0	1	0	ε_2
0	1	1	$\varepsilon_2\varepsilon_3\varepsilon_{23}$	0	0	1	ε_3

This makes clear that the value of the sum depends on the value of the ε-parameters and on s, but not on any specific response pattern that leads to the testlet score of s, whence it follows that the second factor in the right-hand side of (15) is a function of the ε-parameters and the score s. Since it is a sum of exponentials, it is positive, and therefore we can write it as $\exp(-\eta_s(\varepsilon^*))$ or $\exp(-\eta_s)$ for short. Moreover, it is clear from (13) that $\eta_0 = 0$.With this notation, (15) can be written as

$$P(S = s|\theta) \propto \exp(s\theta - \eta_s), \qquad (16)$$

which is formally equivalent to the PCM.

This result is summarized as

Theorem 2 *For any value of the ε^*-parameters in the dependence model (12), and for all testlets consisting of m binary items, there exists a set of m functions $\eta_{1,} \dots, \eta_m$ such that the distribution of the testlet score S in the dependence model is identical to its distribution under the PCM with parameter values $\eta_{1,} \dots, \eta_m$. These functions are given by*

$$\eta_s = -\ln \Gamma_s(\varepsilon^*), \ (s = 1, \dots, m),$$

where $\Gamma_s(\varepsilon^*)$ is defined by (14).

The number of elements in ε^* is $\sum_{g=1}^m |I_g| = 2^m - 1$, so that the parameter space of the dependence model (with the ε-parameterization) is $\mathbb{R}_+^{2^m-1}$. What the theorem says is that the functions $(\eta_1, \dots, \eta_m)$ considered jointly define a vector-valued function from $\mathbb{R}_+^{2^m-1}$ *into* $\mathbb{R}^m$, the parameter space of the PCM at the testlet level. In Figure 1, this result is displayed graphically. The left-hand ellipse represents the parameter space of the dependence model and a dot represents an ε^*-vector. For each such vector there is a (unique) vector in the parameter space of the PCM (right-hand ellipse) representing the equivalent model (at the testlet score level) in the PCM-family.

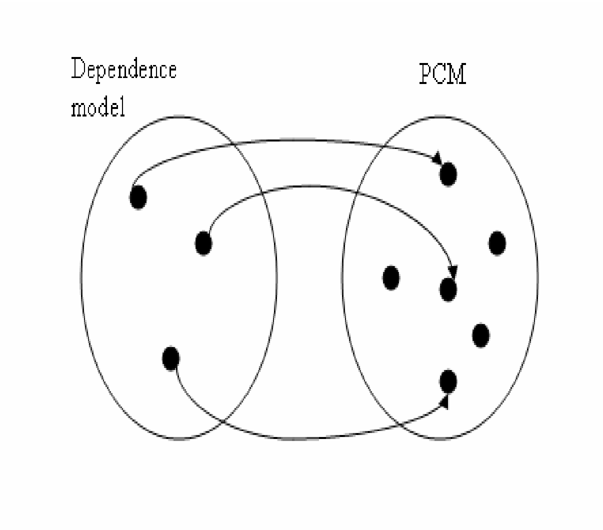

Figure 1: The relationship between the parameter space of the dependence model and the PCM.

This is the main result of this paper: a fairly complicated model for binary responses (the model defined by (12)) can be fitted by using the PCM at the level of testlets. The number of parameters η_s to be estimated is the same as in the Rasch model, but the assumptions are far weaker: complicated patterns of item dependency within testlets are automatically absorbed in the PCM-parameters η_s. Moreover, the sufficient statistic for the latent variable, the raw score, is the same as in the Rasch model.

4.2 The PCM for testlets as a model for sums of binary scores

There remains, however, a complementary problem, which can be seen from Figure 1: in the right-hand ellipse (the parameter space of the PCM) there are dots which are not at the end-point of an arrow, symbolizing vectors in the parameter space of the PCM which cannot be written as the η-transformation of any ε^*-vector in the parameter space of the dependence model. The question to be answered is whether such η-vectors can exist. If they cannot, then we have the result that every partial credit score in the PCM can be interpreted as a sum of m binary item scores, where the distribution of these binary scores is given by the dependence model (12). In the remaining part of this section, it is shown that this is indeed the case.

Since the second factor in the right-hand side of (13) defies simplification, a number of restrictions on the β-parameters will be introduced which yield a more tractable expression, and yet result in a model which covers the parameter space of the PCM. Specifically, we will assume all interaction parameters of the same order to be equal, i.e.,

$$\beta_h = \lambda_g \text{ for all } h \in I_g, \quad (g = 2, \ldots, m). \tag{17}$$

Formally, by applying these restrictions we consider a subspace of the orginal parameter space of the dependence model. Where the original subspace has dimension $2^m - 1$, the restricted subspace has dimension $2m - 1$, because there are m β-parameters with a single subscript and $m - 1$ interaction parameters, $\lambda_2, \ldots, \lambda_m$.

Using the restrictions (17) and the fact that all g-fold products $z_{i_1} \times \cdots \times z_{i_g}$ vanish if $g > s(y)$, and equal one in $\binom{s}{g}$ cases if $g \leq s(y)$, (12) can be rewritten as

$$P(\mathbf{Y} = \mathbf{y}|\theta) \propto \exp\left[s\theta + \sum_{i \in I_1} y_i\beta_i + \sum_{g=2}^{s} \binom{s}{g}\lambda_g\right], \tag{18}$$

whence it follows that (13) simplifies to

$$\begin{aligned} P(S = s|\theta) &\propto \exp(s\theta) \times \exp\left[\sum_{g=2}^{s} \binom{s}{g}\lambda_g\right] \times \sum_{\Sigma z = s} \prod_i \varepsilon_i^{z_i} \\ &= \exp(s\theta) \times \exp\left[\sum_{g=2}^{s} \binom{s}{g}\lambda_g\right] \times \gamma_s(\varepsilon). \end{aligned} \tag{19}$$

Define

$$\eta_s = -\ln \gamma_s(\varepsilon) - \sum_{g=2}^{s} \binom{s}{g} \lambda_g, \quad (s = 1, \ldots, m), \tag{20}$$

where the sum in the right-hand side of (20) is defined to be zero if $s < 2$. Now it is easy to show that for any ordered set of m η-values it is always possible to find ε- and λ-values such that (20) is fulfilled. The values for the ε-parameters can be taken arbitrarily from the positive reals, with the only restriction that minus the logarithm of their sum equals η_1. In this way (20) is fulfilled for $s = 1$. The λ-values are given by sequentially applying (from (20)):

$$\lambda_s = -\ln \gamma_s(\varepsilon) - \eta_s - \sum_{g=2}^{s-1} \binom{s}{g} \lambda_g, \quad (s = 2, \ldots, m). \tag{21}$$

We illustrate this by a simple example for $m = 2$. Suppose $\eta_1 = 0$ and $\eta_2 = 2$. Consider the following two ε-vectors: $\varepsilon^{(1)} = (0.7, 0.3)$ and $\varepsilon^{(2)} = (0.9, 0.1)$. It holds that $\gamma_1(\varepsilon^{(1)}) = \gamma_1(\varepsilon^{(2)}) = 1$, complying in both cases to the restriction that $\eta_1 = -\ln \gamma_1(\varepsilon)$. The basic symmetric functions of order 2, however are not equal in both cases as $\gamma_2(\varepsilon^{(1)}) = 0.21$ and $\gamma_2(\varepsilon^{(2)}) = 0.09$. Applying (21) in both cases, we find

$$\lambda_2^{(1)} = -\ln(0.21) - 2 = -0.439 \text{ and } \lambda_2^{(2)} = -\ln(0.09) - 2 = +0.408$$

and therefore, the two ε^*-vectors $(0.7, 0.3, \exp(-0.439))$ and $(0.9, 0.1, \exp(0.408))$ are transformed into the same η-vector $(0, 2)$. This result is stated formally as

Theorem 3 *The m-valued function $(\eta_1, \ldots, \eta_m)$ defined by (14) over a subspace of the parameter space of the dependence model, defined by (17), is a function from $\mathbb{R}_+^{2m-1}$* onto $\mathbb{R}^m$.

The meaning of this theorem is graphically displayed in Figure 2. The restricted subspace is symbolized by the area in the left-hand ellipse to the right of the waved line. From Theorem 2, we know that there exist an arrow from all points in this subspace to a unique point in the parameter space of the PCM. In Theorem 3, it is stated that *all* points in the parameter space of the PCM are the endpoints of such an arrow. Since $2m - 1 > m$ if $m \geq 2$, this function cannot be one-one; therefore more than one arrow ends in every point of the PCM space.

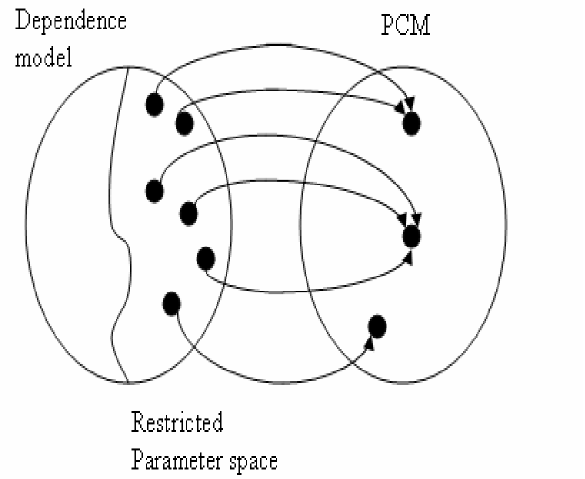

Figure 2: The relationship between the parameter space of the restricted dependence model and the PCM.

In summary, it has been shown that every model in the family defined by (12) is formally equivalent to the PCM when the distribution of the testlet score is modeled (Theorem 2), and conversely, that every PCM can be understood as a model for the testlet score, where the joint distribution of the item responses within the testlet is given by (12) (Theorem 3). If the item responses are observed, then (12) is identified and the parameters may be estimated; if only sums of item scores are observed, however, model (13) results, and the model is no longer identifiable, because there are more parameters than different values of the score. Only functions of these parameters are estimable, for example, the functions given by (14) and one-one transformations of these functions.

The practical implication of this result is discussed in the next section.

5 Practical implications

In applications of the Rasch model, one can focus on different aspects, either paying attention to the structure of the model itself, or focusing on its application, i.e. on the inferences one can make on the latent abilities of concrete persons or groups of persons.

An example of the first is the research with the so-called *Linear Logistic Test Model (LLTM)* (see for example Fischer, 1995; Bechger, Verstralen & Verhelst, 2002), where the item parameters are considered as linear combinations of a (small) number of so-called basic parameters. In these models local independence between item responses is an essential part of the model, and estimates of the parameter values require that data are availble at item level. Detecting that the assumption of local independence is violated in a concrete application of the LLTM invalidates the model immediately, and the results obtained in the previous section cannot be put at use.

There exist, however, other applications where the use of IRT serves a more practical purpose. We take a survey, like national or international assessment in education as a typical situation. There the focus is on the distribution of the target latent variable (e.g., reading literacy) in populations and subpopulations, for example, the comparison of the literacy distributions of boys and girls, in subpopulations that vary in socio-econic status, across different countries and over time. The practical value of using an IRT-approach is that it allows to include much more item material than can be responded to by a single testee, and that it allows to include new item material over time, and at the same time guarantee invariance of the measured concept, although new and old material may differ in difficulty. A large scale project where the Rasch model has been used as IRT model is the PISA project (Adams, 2002). The practical advantage of the results reported in the preceding section is that it does not matter whether the assumption of local independence holds or does not hold, as long as such dependencies are correctly modelled. Applying the PCM at the testlet level is an easy way to capture arbitrary dependencies between items of a testlet, without the necessity of unraveling and testing the precise nature and extent of such dependencies.

Two questions, however, remain to be answered. The first concerns the possible loss in information when one models testlet scores instead of item scores. The second has to do with the vagueness of the notion of testlet in the preceding section. The results were shown to be valid independently of the way the testlets were defined, as long as the testlets were disjoint and the testlet scores locally independent, but it is not not a trivial problem to form such a collection of testlets in a practical application. These two problems will be discussed in turn.

5.1 Loss of information

One might be worried that, if the Rasch model holds, the use of the PCM at the testlet level will lead to information loss, i.e., that the accuracy of the latent variable estimates (or its distribution) will be weaker when based on the PCM rather than on the (correct) Rasch model. There is, however, no reason for such a worry. Both the Rasch model and the PCM are an exponential family of models, and for such models it holds that the Fisher information equals the variance of the sufficient statistic (Barndorf-Nielsen, 1978). The commonly used estimate for the standard error of the θ-estimate is the square root of one divided by the information. In both models, the sufficient statistic for θ is the sum of the testlet scores, and from a comparison of (5) and (8), we see that the distribution of the sufficient statistic for any value of θ is the same in both models, and therefore the variance is the same as well. In case the Rasch model is valid, the PCM is just a reparameterization of the

Rasch model, defined by (7), and the standard errors of the θ-estimates are identical under both models.

But what if the Rasch model is not valid? If the dependence model (14) is valid, but not the Rasch model, then the Fisher information can be determined correctly from it. Of course, the parameters must be estimated from a finite set of data, such that one will obtain only an estimate of the Fisher information. This estimate, however is consistent. If one estimates the variance of the scores using the incorrect Rasch model, the result cannot be interpreted as the Fisher information since the measurement model is not valid, so that comparisons with the Fisher information under the PCM are meaningless.

A related, but quite different question is whether tests with dependent items lead to more of less accuracy of the θ-estimates than tests that comply to the Rasch-model. The answer to this question is not simple, as is shown by the following illustration. Suppose $m = k = 2$ and the parameters β_1 and β_2 are both equal to zero. Now consider three models with these parameters fixed, and the interaction parameter β_{12} taking the values 0, -0.5 and $+0.5$ respectively, as examples of the Rasch model, a dependence model with negative and a dependence model with positive first order interaction respectively. The information functions of these three models are displayed graphically in Figure 3.

The information function for the Rasch model (the solid curve) shows a well-known characteristic of all IRT-models: the accuracy with which θ can be estimated depends on the value of θ itself. In Figure 3, we see that most information is conveyed for $\theta = \beta_1 = \beta_2$. For the dependence models, two characteristics are important, and have shown to be stable for a wide range of parameter values for which similar figures have been scrutinized.

The first is the maximum information of the model. The maxima are located at different places, and it appears that the lower value the of the interaction parameter, the higher the location of maximum information.The maximal information itself, however, seems to correlate

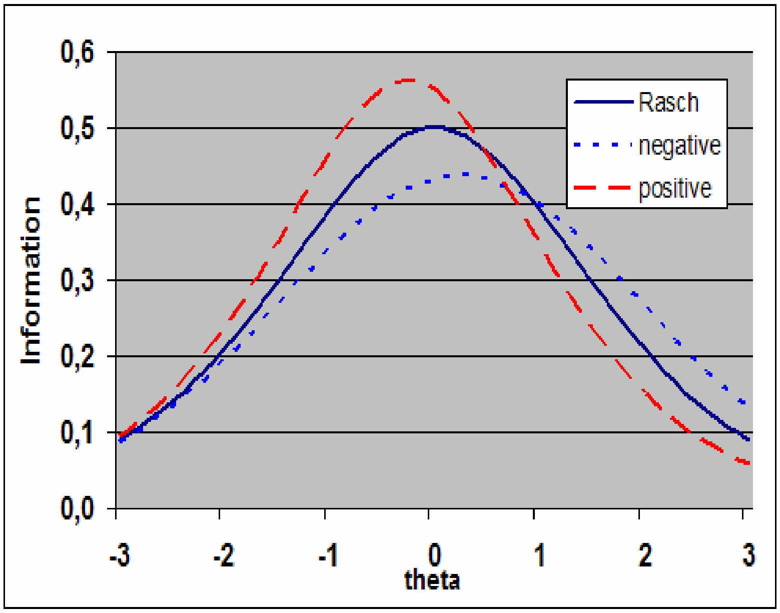

Figure 3: Information functions for a two item test with zero, positive and negative interaction.

positively with the interaction parameter β_{12}: the larger this parameter, the larger the maximal information.

The second characteristic is that all pairs of curves in the figure do intersect. This means that for no model the information is uniformly higher of lower than that of another model. For example, the model in Figure 3 with the lowest modal information ($\beta_{12} = -0.5$) has higher information than the other two for $\theta > 1$.

With more items in a testlet, with more than one testlet and more complicated interactions, it might be far more difficult to describe in general terms the effect of interactions on the information function.

5.2 Detecting interactions

In practical applications, it may not always be easy to detect sets of items where dependence is likely to occur. The most likely candidates are items formulated as questions about the same stem, as is often the case in reading tests. But other dependencies may occur as well, for example in cases where the presence of an item, item i, say, in a linear test contains clues for the solution of another item j. Two methods are discussed to find out whether dependencies are present or not.

The first one departs from a Rasch analysis, where independence is assumed. If conditional maximum likelihood (CML) is used as estimation method, it is fairly simple to construct the matrix of predicted pairwise frequencies of correct responses. The expression is

$$E(n_{ij}) = \sum_{s=2}^{k-1} n_s \frac{\varepsilon_i \varepsilon_j \gamma_{s-2}^{(i,j)}(\varepsilon)}{\gamma_s(\varepsilon)}, \tag{22}$$

where n_s is the frequency of score s in the sample, $\gamma_s(\varepsilon)$ is the gamma function of order s evaluated at the CML-estimates, and $\gamma_{s-2}^{(i,j)}(\varepsilon)$ is the gamma function of order $s-2$, evaluated on the vector of ε-parameters, where ε_i and ε_j are excluded. Response patterns with a score of zero or one are not counted because for these it is impossible to have both items correct, and score k is excluded because the probability of having items i and j correct trivially equals one. Simple or weighted comparison between observed and expected pairwise frequencies may reveal pairs of items where the covariation is too high or too low to be compatible with the assumption of independence. A suitable weighted comparison is

$$z_{ij} = \pm\sqrt{\frac{n^* \left[n_{ij} - E(n_{ij})\right]^2}{E(n_{ij})[n^* - E(n_{ij})]}}, \tag{23}$$

where the sign is the same as the sign of the difference in the numerator of (23), and $n^* = \sum_{s=2}^{k-1} n_s$. The quantity z_{ij}^2 is readily recognized as the common chi-square statistic computed on a 2×1 contingency table with observed frequencies n_{ij} and $n^* - n_{ij}$ respectively. Its signed square root is approximately standard normally distributed.

The second method starts from a PCM analysis and can help in deciding whether the scores on a testlet with maximum score m can be conceived as a sum of m Rasch items. One can proceed along the following lines:

1. Using (7), the PCM parameter estimates can be converted to the coefficients of the m-th degree polynomial P_m given by (10). Using a solution finder, one can find all roots of P_m. If they are all real (and negative by necessity), the Rasch estimates of the parameters ε are given by minus the roots.

2. If not all roots are real, this may be caused by genuine dependencies, but also by sampling error. So we might wish to have a statistical test that enables us to reject the latter hypothesis. It appears to be quite hard to construct such a test, and we did not find a solution to this problem. We can, however, construct a more conservative test, by using Theorem 1 and (7). The null hypothesis, i.e., the Rasch model, can be written in the following two equivalent forms

$$H_0: \frac{(s+1)\,(m-s+1)}{s(m-s)} \times \frac{\gamma_{s-1}(\varepsilon)\gamma_{s+1}(\varepsilon)}{\gamma_s^2(\varepsilon)} \leq 1, \quad (s = 1, \ldots, m-1),$$

or

$$H_0: d_s \equiv 2\eta_s - \eta_{s-1} - \eta_{s+1} + \ln \frac{(s+1)\,(m-s+1)}{s(m-s)} \leq 0, \quad (s = 1, \ldots, m-1). \tag{24}$$

The Wald test statistics are

$$W_s = \frac{\widehat{d}_s^2}{\mathbf{t}'\widehat{\Sigma}_s\mathbf{t}}, \quad (s = 1, \ldots, m-1). \tag{25}$$

where $\widehat{d}_s$ equals d_s evaluated at the ML-estimates, $\widehat{\Sigma}_s$ is the estimated variance-covariance matrix of $\widehat{\eta}_{s-1}$, $\widehat{\eta}_s$ and $\widehat{\eta}_{s+1}$ (in that order) and $\mathbf{t}' = (-1, 2, -1)$. W_s is asymptotically chi-square distributed with one degree of freedom, and therefore its signed square root is standard normally distributed. The sign of the square root is the sign of $\widehat{d}_s$. If $s = 1$, the first row and column of $\widehat{\Sigma}_s$ consist of zeros, since $\widehat{\eta}_0 = \eta_0 = 0$. The null hypothesis is rejected at the 5% level of significance if $W_s > 1.96^2$ and $\widehat{d}_s > 0$.

6 Discussion

In this section, the results of the preceding sections are summarized and some comments are added.

1. The partial credit model is not suited to describe difficulties of item steps. In complex items, where steps can be distinguished, there is no invariant relation between parameter values and the difficulty of the steps. This means that the set of parameters associated with a partial credit item should be considered as a joint formal description of the item as a whole.

2. If the Rasch model holds for a set of k items, the PCM also holds for every partition of the original k item scores in T sum scores defined on T testlets (subsets of items) of arbitrary size. T is arbitrary too. Moreover, there exists a well specified non-linear relationship between the Rasch model parameters and the PCM parameters, given by (7). Although the Rasch parameters can be recovered uniquely from the PCM parameters, it is impossible to associate these values to particular Rasch items, because any permutation of the Rasch parameters of the testlets leads to the same likelihood.

3. One should be careful not to confuse the algebraic equivalence of two models with relations between parameter estimates. We give two comments in this respect.

- Suppose the Rasch model holds, and $m = 2$ for some testlet. Then it follows from Newton's theorem that for the testlet it holds that $\eta_2 \geq 2\eta_1 + \ln 4$. But if one estimates the parameters η_1 and η_2 from a finite data set, even if it is known to comply with the Rasch model, as with artificially generated data, there is nothing that guarantees that this inequality is fulfilled with the estimates. The only thing that is known for sure is that the probability that the inequality is violated goes to zero as the sample size increases without bound. Therefore the maximum of the likelihood function using the PCM at the testlet level will never be smaller than the maximum using the Rasch model. To decide whether the assumption of local independence is credible, one will have to use a statistical test procedure like the one proposed in Section 5.2.
- Although the results discussed are also valid (at the algebraical level) in case $T = 1$, this case cannot be tested empirically, because CML-estimates in the PCM do not exist if the test is composed of one partial credit item.

4. In Section 4, a model for binary items is presented that allows for complicated dependencies between item responses. If such dependencies are restricted to subsets of m items, it is shown that such a model is equivalent to the PCM if testlet scores are modelled instead of binary reponses. Moreover it is shown that each PCM model may be interpreted in this way. This does not imply, however, that such an interpretation also has substantive meaning. The general model (12) is overparameterized if only testlet scores are observed, and an interpretation in terms of these many parameters is a possibility, but certainly not the only one.

5. The practical use of the results mainly resides in the possibility to ignore complicated dependencies between item responses with-

out loosing information about the underlying latent variable. Two methods have been proposed to detect such dependencies, such that the testlet definitions may be adequately chosen.

To conclude, we add a warning against overoptimism. Even if one would succeed completely in identifying subsets of binary items such that the resulting testlet scores are locally independent, this does not imply that the PCM at the testlet score level is the correct model. More general models like the generalized PCM, allowing for different discriminations of testlets, or multidimensional models, or even totally different approaches might point to weaknesses in the simple PCM. There is plenty of room for sustained theoretical research.

References

Adams, R. (2002). Scaling Pisa cognitive data. In R. Adams & M. Wu (Eds.), *PISA 2000 Technical Report*: Paris: OECD.

Andersen, E. (1977). Sufficient statistics and latent trait models. *Psychometrika*, 42, 69–81.

Barndorff-Nielsen, O. (1978). *Information and exponential families in statistical theory.* New York: Wiley.

Bechger, T., Verstralen, H., & Verhelst, N. (2002). Equivalent linear logistic test models. *Psychometrika*, 67, 123–136.

de Vries, H. (1988). *Het partial credit model en het sequentiële Rasch-model met stochastisch design [The partial credit model and the sequential Rasch model with stochastic design].* Technical report, Amsterdam: University of Amsterdam.

Fischer, G. (1995). The linear logistic test model. In G. Fischer & I. Molenaar (Eds.), *Rasch Models: Foundations, Recent Developments and Applications*: New York: Springer Verlag.

Hardy, G., Littlewood, J., & Pólya, G. (1952). *Inequalities.* Cambridge: Cambridge University Press.

Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149–174.

Molenaar, I. (1983). *Item steps.* Technical Report Heymans Bulletin HB-83-630-EX, Groningen: University of Groningen.

Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology*, 43, 39–55.

Tutz, G. (1997). Sequential models for ordered responses. In W. van der Linden & R. Hambleton (Eds.), *Handbook of Modern Item Response Theory*: New York: Springer Verlag.

Van Engelenburg, G. (1997). *On psychometric models for polytomous items with ordered categories within the framework of item response theory.* PhD thesis, University of Amsterdam.

Verhelst, N. & Glas, C. (1995). Dynamic generalizations of the Rasch model. In G. Fischer & I. Molenaar (Eds.), *Rasch Models: Foundations, Recent Developments and Applications*: New York: Springer Verlag.

Verhelst, N., Glas, C., & de Vries, H. (1997). A steps model to analyze partial credit. In W. van der Linden & R. Hambleton (Eds.), *Handbook of Modern Item Response Theory*: New York: Springer Verlag.