Multidimensional Rasch Models for Partial-Credit Scoring

Henk Kelderman

Vrije Universiteit, Amsterdam

Rasch models for partial-credit scoring are discussed and a multidimensional version of the model is formulated. A model may be specified in which consecutive item responses depend on an underlying latent trait. In the multidimensional partial-credit model, different responses may be explained by different latent traits. Data from van Kuyk's (1988) size concept test and the Raven Progressive Matrices test were analyzed. Maximum likelihood estimation and goodness-of-fit testing are discussed and applied to these datasets. Goodness-of-fit statistics show that for both tests, multidimensional partial-credit models were more appropriate than the unidimensional partial-credit model. *Index terms: χ^2 testing, exponential family model, multidimensional item response theory, multidimensional Rasch model, partial-credit models, Progressive Matrices test, Rasch model.*

Responses to educational and psychological test questions can be scored partially correct rather than simply correct or incorrect. To relate a person's responses to the person's underlying latent trait, item response models for dichotomously scored data (Birnbaum, 1968; Lord, 1980; Lord & Novick, 1968; Rasch, 1960/80) have been generalized to polytomous ordered data (Andrich, 1978a, 1978b; Glas & Verhelst, 1989; Masters, 1982; Muraki, 1990; Rost, 1988; Samejima, 1969). In some applications, however, it is questionable whether fully correct answers require the same trait as partially correct answers. A Rasch-type multidimensional partial-credit model (PCM) was formulated for data in which different answers depend on different traits.

Partial-Credit Models

A Unidimensional Model

For the unidimensional PCM (1PCM; Andrich, 1978a; Masters, 1982), let π_{ijx} be the probability of response x of person j to item i and r_i be the number of possible (partial) credit responses for item i. The 1PCM in terms of the log-odds of consecutive item responses can be written as

$$\Omega_{ijx} = \log\left(\Pi_{ijx}/\Pi_{ij(x-1)}\right) = \theta_j - \delta_{ix}, \quad j = 1, \ldots, N, \quad i = 1, \ldots, n, \quad x = 1, \ldots, r_i. \tag{1}$$

with

$$\sum_i \sum_x \delta_{ix} = 0 \tag{2}$$

to fix the scale, and log denotes the logarithm to the base e = 2.718281. That is, the log-odds of consecutive responses can be written as the difference of a person parameter θ_j and a response parameter δ_{ix}. The person parameter describes the trait level of the person. As the person parameter becomes higher, the log-odds of giving a higher quality response x rather than a lower quality response $x - 1$ becomes higher. The threshold parameter describes the difficulty of the item response. As the threshold parameter becomes higher, the log-odds of giving the response x rather than the response $x - 1$ becomes lower. A similar model for the log-odds of nonconsecutive responses is

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 20, No. 2, June 1996, pp. 155–168

0146-6216/96/0200155-14$1.95

$$\log\left(\Pi_{ij2}/\Pi_{ij0}\right) = 2\theta_j - \left(\delta_{i1} + \delta_{i2}\right). \tag{3}$$

This model is a generalization of the Rasch model for dichotomously scored ($r_i = 1$) items. It has the useful property that objects—items or persons—can be compared independently of other objects within the frame of reference (Rasch, 1961, 1977). This property is called specific objectivity (Fischer, 1995; Masters, 1987). In Equation 1, specific objectivity means that two different persons, say j and j', can be compared by

$$\Omega_{ijx} - \Omega_{ij'x} = \theta_j - \theta_{j'} \tag{4}$$

independently of the response x to item i (i,x). Two different responses [e.g., (i,x) and (i',x')] can be compared by

$$\Omega_{ijx} - \Omega_{i'jx'} = \delta_{i'x'} - \delta_{ix} \tag{5}$$

regardless of the person j.

Other unique measurement properties of the Rasch model are that the responses of all items have identical (loglinear) interactions with all other variables in the nomological network (Kelderman, 1995) and that the items follow composite transitivity (Roskam & Jansen, 1984).

Rasch models describe the trait (θ) and difficulty (δ) parameters of individual persons and items, respectively (Rasch, 1960). Not considered in these models is the statistical generalization to other persons from the same population of intended respondents, nor to other items from the same universe of intended items. Therefore, no sampling distributions are specified for person or item parameters. See Molenaar (1995) for a discussion of the different sources of randomness in item response theory (IRT) models.

Multidimensional Partial-Credit Models

The 1PCM with trait parameter θ_j can be generalized to a multidimensional PCM (MPCM) with s trait parameters θ_{jq} $(q = 1, ..., s)$. Let w_{qix} be an indicator variable that takes the value 1 if the log-odds Ω_{ijx} of giving response x rather than $x - 1$ depends on latent trait θ_{jq} and 0 otherwise. The MPCM is then

$$\Omega_{ijx} = \sum_{q=1}^{s} w_{qix}\,\theta_{jq} - \delta_{ix}\,. \tag{6}$$

In this model, the person's ability to give a response to a particular item is considered to be equal to a weighted sum of more basic trait parameters. In the general case of this MPCM, the weights w_{qix} may take discrete non-negative values (0, 1, 2, ...).

Like the 1PCM, this MPCM has the useful property that response alternatives can be compared independently of the person, and persons can be compared independently of the alternatives, provided that both responses depend on the trait of interest.

The model can be written in terms of the response probabilities π_{ijx} rather than the log-odds Ω_{ijx}. Next, the probability of a response pattern and of the entire data matrix are provided.

From Equation 1,

$$\Pi_{ijx} = \exp\left(\sum_{y=1}^{x}\Omega_{ijy}\right)\Pi_{ij0}, \quad x = 1, \ldots, r_i\,. \tag{7}$$

Furthermore from

$$1 = \sum_{x=0}^{r_i}\Pi_{ijx} = \Pi_{ij0} + \sum_{x=1}^{r_i}\Pi_{ijx} = \Pi_{ij0} + \Pi_{ij0}\sum_{x=1}^{r_i}\exp\left(\sum_{y=1}^{x}\Omega_{ijy}\right), \tag{8}$$

$$\Pi_{ijo} = 1 \Big/ \left[1 + \sum_{x=1}^{r_i} \exp\left(\sum_{y=1}^{x} \Omega_{iiy} \right) \right], \tag{9}$$

so that from Equations 6–9

$$\Pi_{ijx} = \frac{\exp\left[\sum_{y=1}^{x} \left(\sum_{q=1}^{s} w_{qiy} \theta_{jq} - \delta_{iy} \right) \right]}{1 + \sum_{z=1}^{r_i} \exp\left[\sum_{y=1}^{z} \left(\sum_{q=1}^{s} w_{qiy} \theta_{jq} - \delta_{iy} \right) \right]}, \tag{10}$$

where

$$\sum_{y=1}^{0} (\cdot) \equiv 0. \tag{11}$$

Note that if $s = 1$ and $w_{qiy} = 1$, the IPCM results.

If the consecutive odds interpretation of the model is not appropriate for the data, an alternative parameterization of Equation 10 is

$$\Pi_{ijx} = \frac{\exp\left(\sum_{q=1}^{s} B_{qix} \theta_{jq} + \phi_{ix} \right)}{1 + \sum_{y=1}^{r_i} \exp\left(\sum_{q=1}^{s} B_{qiy} \theta_{jq} + \phi_{iy} \right)}, \tag{12}$$

where

$$\phi_{i0} \equiv 0, \tag{13}$$

$$\phi_{ix} = \sum_{y=1}^{x} \delta_{iy} \quad (x = 1, \ldots, r_i), \tag{14}$$

and

$$B_{qi0} \equiv 0, \tag{15}$$

$$B_{qix} \equiv \sum_{y=1}^{x} w_{qiy} \quad (x = 1, \ldots, r_i), \tag{16}$$

and conversely,

$$\delta_{ix} \equiv \phi_{i(x-1)} - \phi_{ix} \quad (x = 1, \ldots, r_i), \tag{17}$$

and

$$w_{qix} \equiv B_{qix} - B_{qi(x-1)} \quad (x = 1, \ldots, r_i). \tag{18}$$

This model is a generalization of Rasch's original multidimensional model (Andersen, 1973b; discussed below); it also is a confirmatory multidimensional IRT model.

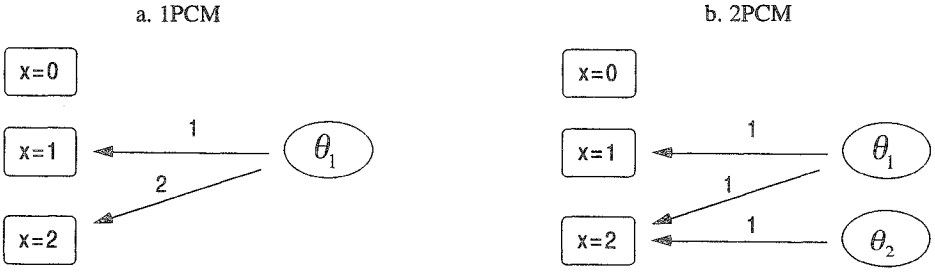
The model is multidimensional because the $\theta_j = (\theta_{j1}, \theta_{j2}, \ldots, \theta_{js})$ are vector valued so that the trait of each person can be represented by a point in an s-dimensional real latent space (Lord & Novick, 1968, p. 359). Note that this multidimensional Rasch model (the MPCM) does not specify a population distribution for θ_j

because the model is concerned with individual characteristics rather than population characteristics.

The model is confirmatory because the B weights, indicating the strength of the relation between the item response and θ_j, must be specified. The model describes the relationship between θ_{jq} and the probability of each of the item responses. If a person's value on θ_{jq} becomes higher, the probability of responses with larger B weights on that trait become higher. Conversely, if a person has many item responses that have high B weights for a certain latent trait, the model states that these responses have a high probability under a high value of this latent trait, so that these responses indicate a high value of the trait.

To specify an ordinary 1PCM, $w_{qix} = 1$ $(q = 1; i = 1, ..., n; x = 1, ..., r_i)$, which is equivalent to specifying $B_{qix} = x$, $(q = 1; i = 1, ..., n; x = 0, ..., r_i)$. Figure 1a describes the influence in the 1PCM of the latent trait θ_1 on each of the responses. The B weights over the arrows describe the strength of the relation. In this model, the correct response $(x = 2)$ depends 2 times more on θ_1 than the partially correct response $(x = 1)$. Figure 1b shows the same situation for a 2PCM. The partially correct response and the fully correct response both depend on θ_1. In addition, the fully correct response also depends on θ_2.

Figure 1
Models for Three-Category Items



Estimating the Multidimensional Partial-Credit Model

The likelihood of the data under the MPCM. Denote the response x of person j to item i as x_{ij} and let $\mathbf{x}_j = (x_{j1}, ..., x_{jn})$ be the vector of responses of person j. Let

$$t_{jq} = \sum_{i}^{n}\sum_{y=1}^{x_{ij}} w_{qiy} \left(= \sum_{i}^{n} B_{qix_{ij}} \right) \tag{19}$$

be person j's sum of weights on trait q; $\mathbf{t}_j = (t_{j1}, ..., t_{js})$ be the vector of weight-sums; and $\boldsymbol{\delta}$ be the vector of threshold parameters. If it is assumed that the person's responses are independent, the probability of this response vector can be derived from Equations 6 and 7 as

$$p\left(\mathbf{x}_j \middle| \boldsymbol{\theta}_j\right) = \prod_{i=1}^{n} \Pi_{ijx_{ij}} = \exp\left[\sum_{i=1}^{n}\sum_{y=1}^{x_{ij}} \left(\sum_{q=1}^{s} w_{qiy}\theta_{jq} - \delta_{iy} \right) + \sum_{i=1}^{n} \ln\Pi_{ij0} \right] = g\left(\boldsymbol{\delta}, \boldsymbol{\theta}, \mathbf{t}_j\right) h\left(\boldsymbol{\delta}, \mathbf{x}_j\right), \tag{20}$$

where

$$g\left(\boldsymbol{\delta}, \boldsymbol{\theta}_j, \mathbf{t}_j\right) = \exp\left(\sum_{q=1}^{s} \theta_{jq} t_{jq} + \sum_{i=1}^{n} \ln\Pi_{ij0} \right), \tag{21}$$

and

$$h\left(\boldsymbol{\delta}, \mathbf{x}_j\right) = \exp\left(\sum_{i=1}^{n}\sum_{y=1}^{x_{ij}} \delta_{iy} \right). \tag{22}$$

Estimation of parameters. The model in Equation 20 contains n item parameters and N person parameters. It may seem natural to estimate both item parameters and person parameters maximizing the likelihood

$$P(\mathbb{X}) = \prod_{j=1}^{N} p\left(\mathbf{x}_j \middle| \boldsymbol{\theta}_j\right), \tag{23}$$

where $\mathbb{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)'$ is the data matrix. However, as the number of person parameters increases with N, the parameter estimates become inconsistent (Neyman & Scott, 1948). To remedy this situation, a conditional estimation procedure is used.

The model in Equation 20 is an exponential family distribution (Lehman, 1983, p. 26), and $\mathbf{t}_j = (t_{j1}, \ldots, t_{js})$ is a sufficient statistic for the person parameters $\boldsymbol{\theta}_j = (\theta_{j1}, \theta_{j2}, \ldots, \theta_{js})$. This implies that all statistical information in the data about the person's position in the multidimensional latent space $\boldsymbol{\theta}_j$ is contained in the $\mathbf{t}_j$; that is, there is a one-to-one relation between parameter estimates and the statistic. Note that $\mathbf{t}_j$ is a statistic because the model is confirmatory; that is, it depends on the data through weights (B_{qix} or w_{qix}) that are not estimated but specified by the user. Exponential family models allow for conditional inference (Lehman, 1983) as follows.

If $\Sigma_{\mathbf{y}|\mathbf{t}_j}$ is defined as the sum over all possible response vectors $\mathbf{y}$ that give rise to a weight-sum vector $\mathbf{t}_j$, the probability of a vector $\mathbf{t}_j$ can be written as

$$p\left(\mathbf{t}_j \middle| \boldsymbol{\theta}_j\right) = \sum_{\mathbf{y}|\mathbf{t}_j} p\left(\mathbf{y} \middle| \boldsymbol{\theta}_j\right) = g\left(\boldsymbol{\delta}, \boldsymbol{\theta}_j, \mathbf{t}_j\right) \sum_{\mathbf{y}|\mathbf{t}_j} h(\boldsymbol{\delta}, \mathbf{y}). \tag{24}$$

A convenient property of Rasch models is that the conditional probability of $\mathbf{x}_j$ given $\mathbf{t}_j$ does not depend on $\boldsymbol{\theta}_j$:

$$P\left(\mathbf{x}_j \middle| \mathbf{t}_j\right) = P\left(\mathbf{x}_j \middle| \boldsymbol{\theta}_j\right) \Big/ P\left(\mathbf{t}_j \middle| \boldsymbol{\theta}_j\right) = h\left(\boldsymbol{\delta}, \mathbf{x}_j\right) \Big/ \sum_{\mathbf{y}|\mathbf{t}_j} h(\boldsymbol{\delta}, \mathbf{y}) = h\left(\boldsymbol{\delta}, \mathbf{x}_j\right) \Big/ f\left(\boldsymbol{\delta}, \mathbf{t}_j\right), \tag{25}$$

which results from division of Equation 20 by Equation 24. This means that estimation of the threshold parameters $\boldsymbol{\delta}$ can be pursued independently of the estimation of person parameters. Note that this property holds for both multidimensional and unidimensional models, because Equation 20 is an exponential family model.

Let $S_{\mathbf{t}}$ be the number of persons with weight-sum vector $\mathbf{t}$ and let R_{iy} be the number of persons with a response of at least y to item i. Furthermore, let $\mathbb{T} = (\mathbf{t}_1, \ldots, \mathbf{t}_N)'$ be the matrix of weight-sums. The conditional probability of the data given the weight-sums then becomes

$$P(\mathbb{X}|\mathbb{T}) = \prod_{j=1}^{N} p\left(\mathbf{x}_j \middle| \mathbf{t}_j\right) = \exp\left[\sum_{j=1}^{N} \log h\left(\boldsymbol{\delta}, \mathbf{x}_j\right) + \sum_{j=1}^{N} \log f\left(\boldsymbol{\delta}, \mathbf{t}_j\right)\right] = \exp\left[\sum_{i} \sum_{y} \delta_{iy} R_{iy}(\boldsymbol{\delta}, \mathbf{y}) + \sum_{\mathbf{t}} \tau_{\mathbf{t}} S_{\mathbf{t}}\right], \tag{26}$$

where $\tau_{\mathbf{t}} = \log f(\boldsymbol{\delta}, \mathbf{t})$ is a constant of proportionality. Viewed as a function of the parameters $\boldsymbol{\delta}$, this is the conditional likelihood function, which takes its maximum for the maximum likelihood estimator of $\boldsymbol{\delta}$. This likelihood has the form of an exponential family with sufficient statistics R_{iy} for the threshold parameters δ_{iy} and numbers $S_{\mathbf{t}}$ for each of the proportionality constants $\tau_{\mathbf{t}}$ (Lehman, 1983, p. 43). Standard theory for exponential family models yields the likelihood equations $R_{ix} = \mathrm{E}(R_{ix})$ and $S_{\mathbf{t}} = \mathrm{E}(S_{\mathbf{t}})$. Solving these equations for the parameters $\boldsymbol{\delta}$ and τ gives the maximum likelihood estimates (MLEs) of the parameters. The equations, however, have no direct solution so they must be solved by iterative methods. For the case of the PCM ($s = 1$), efficient algorithms have been described by Andersen (1973a, 1977) and Fischer (1974). For the case of a discrete-data-exponential-family likelihood of the form of Equation 26, more general algorithms for the analysis of discrete data (Andersen, 1980; Bishop, Fienberg, & Holland, 1975; Haberman, 1979; Kelderman, 1992) may be used. Appendix A describes an adaptation of a Deming-Stephan (Deming & Stephan, 1940; Bishop et al., 1975, p. 84) algorithm; Appendix B describes the Newton-Raphson algorithm. Both algorithms yield MLEs.

Newton-Raphson is faster, but it may break down if the starting values are far from the MLEs. It also requires the computation of second-order derivatives of the likelihood function. The Deming-Stephan algorithm, however, is not sensitive to the choice of starting values and requires the sufficient statistics R_{ix} and S_t only. Deming-Stephan takes more iterations than Newton-Raphson, but each iteration takes less computation. The Newton-Raphson algorithm was used here after a few initial Deming-Stephan iterations to obtain good starting values. Both algorithms are implemented in the computer program LOGIMO (Kelderman & Steen, 1993).

A complicated problem for both algorithms is the computation of the expected sufficient statistics $E(R_{ix})$ and $E(S_t)$. Efficient solutions for the PCM were described by Andersen (1977) and Fischer (1974). For more general discrete data models, efficient computational procedures have been described by Kelderman (1992).

Testing Multidimensionality of Partial-Credit Models

For exponential family models based on discrete data (Equation 26), various goodness-of-fit statistics are available. Bishop et al. (1975) and Haberman (1979) provided a complete account of these statistics [e.g., Pearson's statistic (X^2) and the likelihood-ratio statistic (G^2)]. These statistics are asymptotically distributed as χ^2s with degrees of freedom (df) equal to the difference between the number of possible response patterns and the number of independent model parameters. If, however, the expected counts of the response patterns become too small, the approximation of the distribution of the overall goodness-of-fit statistic X^2 and G^2 by a χ^2 distribution becomes poor (Koehler, 1977; Lancaster, 1961), although the distribution of X^2 is generally closer to χ^2 than to G^2 (Cox & Placket, 1980; Larnz, 1978). The usual criterion for the size of the expected counts is 5; however, if the distribution is smooth the minimum expected count could be as small as 1 (Cochran, 1952, 1954).

If the minimum expected counts are too small, the model of interest can be tested against an alternative model that contains the PCM of interest as a special case, but that also contains parameters describing a particular deviation. This alternative model will be called a *diagnostic model.* For the size concept data discussed below, a possible diagnostic model for the 2PCM would be a 2PCM with different item response parameters in each age group.

A statistic that can be used to compare a PCM with such a diagnostic model is the likelihood-ratio statistic. Let $L = \log P(\mathbf{X}|T)$ be the loglikelihood of the model of interest and L^* be the loglikelihood of the diagnostic model. The likelihood-ratio test statistic is

$$LR = -2(L - L^*), \tag{27}$$

which is asymptotically distributed as χ^2 with df equal to the difference in the number of linearly independent parameters of both models (Rao, 1973, pp. 418–420). Haberman (1977) showed that the likelihood-ratio statistic has good asymptotic properties.

Example Applications

Example 1: Size Concept Data

Data

Van Kuyk (1988) collected data from an observation program of 4–6.5 year-old children. The program tested skills prerequisite for arithmetic abilities. One subtest was reanalyzed here because it requires partial-credit scoring. The test measured the application of size concepts such as "long-short," "high-low," "thick-thin," and "wide-narrow." For example, in Item 6, a figure displaying four identical girls with successively shorter skirts was shown to the child. The test administrator pointed to the figure and said "Here you see some skirts. They gradually become a bit ..." and the examinee supplies the answer. Answers were rated incorrect if the answer was cognitively incorrect (e.g., "different" rather than "short"). Answers were rated correct if

the correct size concept was given (i.e., "long-short") and was correctly applied [e.g., "short(er)" rather than "long(er)"]. Linguistic errors, such as "small" rather than "smaller," did not count as long as the answer was cognitively correct. Small children may be unable to produce the correct specific concept (e.g., "long-short"), but may use the general size concept "big-small" instead. If "big-small" was correctly applied [e.g., "the skirt is small(er)"] the answer was rated partially correct.

The analysis of these data focused on the question of whether the ability to apply the "big-small" general concept was the same latent trait as the ability to apply the specific dimensional size concept "long-short." It may be hypothesized that these traits are not identical because the first activity may be based solely on the perceptual saliency of a picture element (e.g., the skirt), whereas the second process depends on the correct identification of a particular figural property of that element (e.g., vertical length).

Each of the $N = 263$ persons responded to $n = 15$ items. The random response X_{ij} of fixed person j ($j = 1, ..., N$) on fixed item i ($i = 1, ..., n$) was modeled. In this example, this response was scored with values $x_{ij} = 0$ for an incorrect response, $x_{ij} = 1$ for a partially correct response, and $x_{ij} = 2$ for a correct response, where $x = 1, ..., r_i$.

The sample was divided into three age groups: for ages 4–5 there were 66 children; ages 5–5.5, 132 children; and ages 5.5–6, 65 children. The likelihood equations of the PCM and MPCMs estimated here were solved simultaneously in all three subpopulations, and the item response parameters δ were set equal over subgroups. The τ_t parameters were allowed to vary between subgroups. This required a slight extension of the likelihood equations. To accomplish this, a subgroup subscript—l—was added to S and τ in the likelihood equations: $S_{tl} = E(S_{tl})$.

Results

PCM. Table 1 shows the item threshold parameter estimates ($\hat{\delta}_{ix}$) of the 1PCM for the size concept data. The parameters $\hat{\delta}_{i1}$ pertain to the log-odds of the partially correct response ($x = 1$) relative to the incorrect response ($x = 0$), the parameters $\hat{\delta}_{i2}$ pertain to the log odds of the correct response ($x = 2$) relative to the partially correct response ($x = 1$), and the parameters $\hat{\delta}_{i1} + \hat{\delta}_{i2}$ pertain to the log-odds of the correct response ($x = 2$) relative to the incorrect response ($x = 0$). Note that the identifying constraint in Equation 2 fixes the sum of all parameters $\hat{\delta}_{i1}$ and $\hat{\delta}_{i2}$ to be equal to 0, so that the mean of the parameters $\hat{\delta}_{i1} + \hat{\delta}_{i2}$ is 0.

Table 1 shows that the mean (−.37) of $\hat{\delta}_{i1}$ for the partially correct responses was lower than 0. This means that, on average, it was less difficult to give a "big-small" response than to give a correct specific size concept (e.g., "shorter") relative to an incorrect response. For example, in Item 6 it was relatively easy to see that the rightmost skirt was "smaller" ($\hat{\delta}_{6.1}$) but it was more difficult to apply the correct size concept "shorter" both relative to the incorrect response ($\hat{\delta}_{6,1} + \hat{\delta}_{6.2} = 1.36$) and to the partially correct response ($\hat{\delta}_{6,2} = 1.35$).

Item 8 is an item for which the reverse is true. For Item 8, a figure is shown with three ladders. On each of the ladders is an identical child standing on a rung, but the children are standing on lower positions on consecutive ladders. In Item 8 it is relatively difficult to give the "big-small" answer ($\hat{\delta}_{81} = 2.78$) rather than the incorrect answer, but easier to give the correct answer ($\hat{\delta}_{82} = -1.60$) "lower" rather than the "big-small" answer. This seems rather obvious because in the pictures there are no elements differing in size, only position, whereas in Item 6 the skirts clearly differ in size.

To get the fully correct answer, the specific type of size concept also seemed to matter. Concepts describing the vertical dimension—"low, less, lower," and "higher"—seemed to be more difficult than the size concepts that were not associated with a particular direction such as "long, short, thin, thickest, longest," and "thinnest."

MPCM. A 2PCM was also specified for these data. It was hypothesized that individual differences in the ability to apply the general size concept "big-small" were qualitatively different from the individual differences in the ability to apply the specific size concept (e.g., "long-short"). Therefore, the MPCM was

Table 1
Parameter Estimates of the 1PCM and the 2PCM for the Size Concept Data

Item and Content	1PCM			2PCM Normed as in 1PCM			2PCM Normed to Zero Means		
	δ_{i1}	δ_{i2}	$\delta_{i1}+\delta_{i2}$	δ_{i1}	δ_{i2}	$\delta_{i1}+\delta_{i2}$	δ_{i1}	δ_{i2}	$\delta_{i1}+\delta_{i2}$
1 (Long)	−.79	−1.28	−2.07	−.69	−1.39	−2.09	−.32	−1.76	−2.09
2 (Low)	.59	2.34	2.93	.52	2.40	2.92	.89	2.03	2.92
3 (Short)	−2.14	−.17	−2.31	−2.12	−.19	−2.31	−1.75	−.57	−2.31
4 (A lot)	.30	−1.27	−.97	.42	−1.39	−.97	.79	−1.79	−.97
5 (Thin)	−.83	−.86	−1.69	−.75	−.95	−1.70	−.38	−1.32	−1.70
6 (Shorter)	.01	1.35	1.36	−.03	1.41	1.38	.34	1.04	1.38
7 (Less)	.61	1.33	1.94	.58	1.36	1.94	.95	.99	1.94
8 (Lower)	2.78	−1.60	1.18	2.92	−1.73	1.19	3.29	−2.10	1.19
9 (Thickest)	.13	−.21	−.08	.18	−.25	−.06	.56	−.62	−.06
10 (Higher)	−.85	2.17	1.32	−.97	2.30	1.33	−.60	1.93	1.33
11 (Longest)	−.84	.11	−.73	−.82	.11	−.71	−.45	−.26	−.71
12 (Thickest)	−.98	−.51	−1.49	−.92	−.56	−1.49	−.55	−.93	−1.49
13 (Most)	−1.35	1.98	.63	−1.47	2.12	.65	−1.10	1.75	.65
14 (Thinnest)	−.03	−.89	−.92	.06	−.98	−.92	.43	−1.35	−.92
15 (Highest)	−2.30	3.11	.81	−2.45	3.29	.84	−2.08	2.92	.84
Mean	−.37	.37	0.00*	−.37*	.37*	0.00*	0.00*	0.00*	0.00*

*Fixed.

specified such that $q = 1, 2$ and $x = 0, 1, 2$ with weights $w_{qix} = 1$ if $q = x$ and 0 otherwise.

The item parameter estimates for the 2PCM are also given in Table 1. In the 2PCM, the origin of both latent dimensions was indeterminate. To remove this indeterminacy, means of the item parameters for each latent dimension were set equal to a constant. In Table 1, the parameter estimates of the 2PCM are displayed for two choices of identifying constants (0.00, 0.00) and (−.37, .37) corresponding to the mean values found under the 1PCM. The latter normalization simplifies comparison with the 1PCM. Table 1 shows that the parameter estimates of the 2PCM were approximately the same as in the 1PCM. For example, for Item 15 ($\delta_{15,1}$ and $\delta_{15,2}$) were −2.30 and 3.11 for the 1PCM and −2.45 and 3.29 for the 2PCM, respectively.

Goodness of fit. To select between the models, their goodness of fit was compared. The number of possible response patterns for the size concept data (315) was too large to use the overall X^2 or G^2 goodness-of-fit statistics. Therefore, the likelihood ratio statistic of Equation 27 was used to compare the fit of different models. Four models were considered: two versions of the 1PCM and two versions of the 2PCM. The first version had threshold parameters that were the same across age groups and the second version had threshold parameters that differed across age groups. Table 2 gives −2 times the likelihood and the number of parameters of each of these models. These numbers were used to compute the likelihood ratio statistic of Equation 27. Comparing the fit of the 1PCM with equal (invariant) parameters over subgroups with the 1PCM with different (dependent) parameters over subgroups (Row 1), a large likelihood ratio statistic was obtained (3,436.51 − 3,280.01 = 156.50) relative to the *df* (57). Therefore, the former model was rejected against the latter at the .05 significance level.

Example 2: Progressive Matrices Items

Method

Data. Data from 1,061 persons to the Standard Raven Progressive Matrices Test (Raven, Raven, & Court, 1991) collected were reanalyzed (Vodegel Matzen, 1994). The Progressive Matrix Test is a nonverbal test that is usually assumed to measure the single trait of analytic intelligence.

The fit of several PCMs was compared and the best-fitting model was selected. Because chance capitalization may affect the results, the findings were cross-validated on an independent sample. To do this, the sample

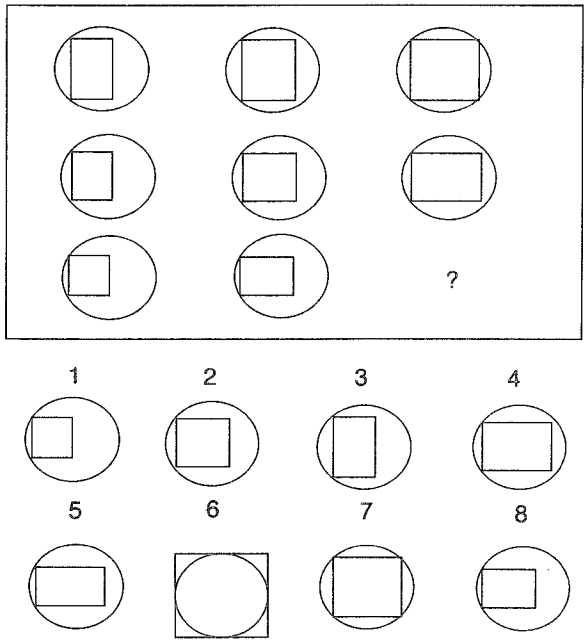
Table 2
−2 Times the Likelihood (−2L) and Number of Parameters (No.) for the 1PCM and the 2PCM

	Subgroup Invariant Parameters		Subgroup Dependent Parameters		Difference		
Model	−2L	No.	−2L	No.	LR	*df*	*p*
1PCM	3,436.51	108	3,280.01	165	156.50	57	0.00
2PCM	3,298.51	152	3,141.92	209	156.59	57	0.00
Difference	138.00	44	138.09	44			

was randomly divided into an exploration sample ($N = 511$) and a cross-validation sample ($N = 550$).

Figure 2 shows a typical matrix item. To protect the security of the actual Raven items, an isomorph (which uses the same rules but different figural elements and attributes) is shown here. The items have eight answer alternatives. The person is asked to select the alternative that fits the matrix best. The answers can be found by scanning certain rules; some items require different rules than others. In the item in Figure 2, the height of the square diminishes in each row and the width of the square increases in each column. Both rules are called quantitative pairwise progression (Rule 1). Alternative 5 is the correct response—it has the largest width and the smallest height. Alternatives 1, 4, 7, and 8 are partially correct responses: either the height is correct and the width is incorrect (1 and 8), or the width is correct and the height is incorrect (4 and 7). The remaining responses (1, 3, and 6) are incorrect responses.

Figure 2
Isomorph of Item C7 of the Raven Progressive Matrices Test



To illustrate the use of MPCMs in studying the dimensionality of matrix items, six items were selected ($n = 6$): three items—C6, C7, and C10—were based on Rule 1 and three items—D7, D8, and D9—used a second rule (Rule 2). Table 3 shows for each response alternative whether it was incorrect (denoted by $x = 0$), partially correct ($x = 1$), or fully correct ($x = 2$).

Table 3
Scoring of Responses to Six Matrix Items Into the Categories Incorrect ($x = 0$), Partially Correct ($x = 1$), and Correct ($x = 2$)

	Response Alternative							
Item	1	2	3	4	5	6	7	8
1 (C6)	1	0	1	2	1	0	0	0
2 (C7)	1	0	0	1	2	0	1	1
3 (C10)	1	1	0	0	0	2	1	1
4 (D7)	0	1	0	1	2	0	1	1
5 (D8)	1	1	1	2	1	0	0	0
6 (D9)	2	1	1	1	1	0	0	0

Models. As multidimensional alternatives to the IPCM (Model A), several MPCMs may be specified for these data. One alternative hypothesis (Model B) is based on the idea that finding two instances of the rule rather than one requires more cognitive ability and not necessarily the same type as finding one instance only. Sternberg (1990, p. 121) discussed control processes that are involved in the execution of compounded cognitive tasks—such as a person keeping track of their place in task performance. It may be hypothesized that finding both instances of the rules invokes such a type of ability on which persons may differ. This hypothesis can be modeled by specifying a separate latent trait pertaining to the fully correct response (see Figure 1b).

Another alternative hypothesis (Model C) is that finding Rule 1 and Rule 2 instances pertain to different cognitive abilities. This hypothesis may be modeled by specifying a different latent trait for Items C6, C7, and C10 than for Items D7, D8, and D9.

Table 4 gives the B weight specifications for Rule 1 items and Rule 2 items of four MPCMs. Model B is two-dimensional within each item; that is, the fully correct response involves an additional latent trait corresponding to a specific metacognitive ability that is necessary to correctly combine two rule instances. Table 4 shows that in Model B both Rule 1 and Rule 2 items pertained to the same latent traits.

Model C is like Model A in that within each item only one latent trait is specified (see also Figure 1a). The difference between A and C is that different latent traits are specified for the items that involve Rule 1

Table 4
B Weight Specifications for MPCMs for Rule 1 Items ($j = 1, 2, 3$) and Rule 2 Items ($j = 4, 5, 6$) for Response Categories Incorrect ($x = 0$), Partially Correct ($x = 1$), and Correct ($x = 2$)

Model,	Rule 1 Items			Rule 2 Items		
θ, and	x			x		
Interpretation	0	1	2	0	1	2
Model A						
θ_1: Find Rules General	0	1	2	0	1	2
Model B						
θ_1: Find Rules General	0	1	1	0	1	1
θ_2: Metacomponent	0	0	1	0	0	1
Model C						
θ_1: Find Rules of Type 1	0	1	2	0	0	0
θ_2: Find Rules of Type 2	0	0	0	0	1	2
Model D						
θ_1: Find Rules of Type 1	0	1	1	0	0	0
θ_2: Find Rules of Type 2	0	0	0	0	1	1
θ_3: Metacomponent	0	0	1	0	0	1

(Items 1, 2, and 3) than for the items that involve Rule 2 (Items 4, 5, and 6). Finally, Model D has both multidimensionality within and across the items. This model has three latent traits: one latent trait corresponding to Rule 1, one corresponding to Rule 2 (as in Model C), and one corresponding to the metacomponent (as in Model B).

Results

To test unidimensionality within the item responses, Model A was compared with Model B and Model C was compared with Model D using a likelihood ratio (LR) test. Both comparisons [LR(A,B) = 102, df = 13; LR(C,D) = 132, df = 8] were significant at the .05 significance level. Thus, the null hypothesis of a 1PCM within each item was rejected in favor of the 2PCM.

To test the hypothesis that Rule 1 items pertained to a different latent trait than Rule 2 items, Model A was compared with Model C and Model B was compared with Model D. These comparisons were also highly significant [LR(A,C) = 147, df = 34; LR(B,D) = 177, df = 29], indicating that both types of items pertained to different latent traits.

Table 5 shows item fit for Model D. In Table 5, Model D's X^2_{iq} statistics are given for each combination of item (i) and θ for which weights were specified in the model. Comparing these χ^2 statistics with their df, no significant outcomes were found that would lead to the rejection of Model D for the exploration sample. Therefore, it was concluded that Model D, specifying a different latent trait for each of the rules and a common latent trait for the metacognitive ability in combining the rules, fit the data. Table 5 also shows the cross-validation results of the X^2_{iq} statistics for Model D. These results also indicated a good fit to the data. In the cross-validation sample, none of the statistics exceeded the $\alpha = .05$ critical value of 11.1 (df = 5) or 19.7 (df = 11).

Table 6 shows the parameter estimates for Model D. For each latent trait, one parameter was fixed to 0 to

Table 5
X_{jq} Goodness-of-Fit Statistics of the Three-Dimensional PCM (Model D) for Progressive Matrices Data for the Exploration Sample and the Cross-Validation Sample

Sample and		Item					
Latent Trait	df	1 (C6)	2 (C7)	3 (C10)	4 (D7)	5 (D8)	6 (D9)
Exploration Sample							
1	5	8.11	3.86	3.32			
2	5				3.83	7.41	1.76
3	11	13.90	7.88	7.20	4.89	12.00	7.73
Cross-Validation Sample							
1	5	8.39	4.89	4.72			
2	5				1.88	1.45	2.59
3	11	6.69	8.68	17.69	8.21	6.83	4.92

make the model identifiable: δ_{11} fixed the scale of the first latent trait (Rule 1), δ_{41} fixed the scale of the second latent trait (Rule 2), and δ_{12} fixed the scale of the third latent trait (metacomponent). The δ parameters for consecutive odds as well as the φ parameters are given. Note that the δs can only be compared if they pertain to the same latent trait. For example, the parameters δ_{i2} ($i = 1, ..., n$) all pertain to the third latent trait and can, therefore, be compared. Table 6 shows that Items 1 and 2 were easier on the third latent trait than Items 3–6 (their parameters $\delta_{12} = 0.00$ and $\delta_{22} = .07$ were smaller than $\delta_{32} = 1.25$, $\delta_{42} = 1.02$, $\delta_{52} = .79$, and $\delta_{62} = 1.24$).

Discussion

By applying MPCMs to van Kuyk's size concept data and the Raven Progressive Matrices Test, it was shown that MPCMs can be used to test hypotheses about the dimensionality of polytomous test data. The

Table 6
Parameter Estimates (ϕ) of the Three-Dimensional PCM (Model D) for Progressive Matrices Data for the Total Sample

Item, δ, and ϕ	Item Score 1	Item Score 2
1 (C6)		
δ	0.00*	0.00*
ϕ	0.00*	0.00*
2 (C7)		
δ	−.27	.07
ϕ	.27	.20
3 (C10)		
δ	1.28	1.25
ϕ	−1.28	−2.53
4 (D7)		
δ	0.00*	1.02
ϕ	0.00*	−1.02
5 (D8)		
δ	−.05	.79
ϕ	.05	−.74
6 (D9)		
δ	.62	1.24
ϕ	−.62	−1.86

*Fixed to 0.

2PCM of the size concept example was a special case of the general MPCM in which all partially correct scores pertained to one latent trait and all fully correct scores pertained to one latent trait. In that case the MPCM is in fact equivalent to Rasch's original multidimensional model (Andersen, 1973b). To show this, let $r_i = s$ and $w_{qiy} = 1$ for $q = y$, and 0 otherwise; then:

$$\ln\left(\Pi_{ijx}/\Pi_{ij0}\right) = \sum_{y=1}^{x}\Omega_{ijy} = \sum_{y=1}^{x}\sum_{q=1}^{s}w_{qiy}\theta_{jq} - \sum_{y=1}^{x}\delta_{iy} = \sum_{y=1}^{x}\theta_{jy} - \sum_{y=1}^{x}\delta_{iy} = \theta_{jx}^{*} - \delta_{ix}^{*}, \tag{28}$$

which is the MPCM. If, however, there are different specifications of w_{qiy}, such as in the Progressive Matrices example, this may no longer be the case.

To test the goodness of fit of the items in a certain MPCM, item by weight-sum goodness-of-fit statistics were used. Further research is needed on how to combine these statistics into one overall goodness-of-fit statistic. For the dichotomous Rasch model, van den Wollenberg (1979, 1982) proposed the Q_1 statistic, which is a weighted sum of the item by sum-score statistics X_{i1}^2 for the dichotomous Rasch model. Although simulations have shown that the distribution of the statistic is close to χ^2, no proof of this is available to date. Glas (1988) derived a similar statistic for which he proved that it is asymptotically a χ^2. The statistic involves the inversion of a second derivative matrix that is more difficult to obtain. It is worthwhile to study a generalization of this statistic for MPCMs.

Appendix A

Kelderman (1992) described Deming-Stephan and Newton-Raphson algorithms to solve the likelihood equations iteratively. Let $K_{ix} = R_{ix} - R_{i(x-1)}$ be the number of persons with response x on item i. Starting with parameter values of 0, the Deming-Stephan algorithm (also called iterative proportional fitting) yields the MLEs by repeated application of

$$\hat{\tau}_t^{(new)} = \hat{\tau}_t^{(old)} + \left[\log S_t - \log \mathrm{E}(S_t)^{(old)}\right], \tag{29}$$

and

$$\hat{\phi}_{ix}^{(new)} = \hat{\phi}_{ix}^{(old)} + \left[\log K_{ix} - \log \mathrm{E}(K_{ix})^{(old)}\right], \tag{30}$$

where "old" and "new" denote the parameter values before and after iteration, respectively. The $\hat{\delta}_{ix}$ parameters can then be obtained from

$$\hat{\delta}_{ix} = \hat{\phi}_{i(x-1)} - \hat{\phi}_{ix} \quad (x = 1, \ldots, r_i). \tag{31}$$

Appendix B

The Newton-Raphson algorithm is faster but more sensitive to the choice of starting values than the Deming-Stephan algorithm. It is based on the iteration

$$\begin{bmatrix} \hat{\boldsymbol{\phi}}^{(new)} \\ \hat{\boldsymbol{\tau}}^{(new)} \end{bmatrix} = \begin{bmatrix} \hat{\boldsymbol{\phi}}^{(old)} \\ \hat{\boldsymbol{\tau}}^{(old)} \end{bmatrix} + \begin{bmatrix} \mathbb{K} - \mathrm{E}(\mathbb{K})^{(old)} \\ \mathbb{S} - \mathrm{E}(\mathbb{S})^{(old)} \end{bmatrix} \left[\mathbb{H}^{(old)}\right]^{-1}, \tag{32}$$

where $\boldsymbol{\phi} = (\phi_{ix})$, $\boldsymbol{\tau} = (\tau_t)$, $\mathbb{K} = (K_{ix})$, $\mathbb{S} = (S_t)$, and $\mathbb{H}$ is the Hessian matrix. Let $\mathbb{M}_{11} = (M_{ix i'x'})$, where $M_{ix i'x'}$ is the number of persons with response x on item i and response x' on item i'. Define $\mathbb{M}_{12} = \mathbb{M}_{21'} = (M_{ixt})$ similarly. Furthermore, let $\mathbb{M}_{22} = \mathrm{diag}(\mathbb{S})$. The Hessian is then

$$\mathbb{H} = \mathrm{E}\left(\begin{bmatrix} \mathbb{M}_{11}\mathbb{M}_{12} \\ \mathbb{M}_{21}\mathbb{M}_{22} \end{bmatrix}\right). \tag{33}$$

References

Andersen, E. B. (1973a). *Conditional inference and models for measuring*. Unpublished doctoral dissertation, Mentalhygiejnisk Forskningsinstitut, Copenhagen.

Andersen, E. B. (1973b). Conditional inference and multiple choice questionnaires. *British Journal of Mathematical and Statistical Psychology, 26,* 31–44.

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

Andersen, E. B. (1980). *Discrete statistical models with social science applications.* Amsterdam: North Holland.

Andrich, D. (1978a). A rating scale formulation for ordered response categories. *Psychometrika, 43,* 561–573.

Andrich, D. (1978b). Scaling attitude items constructed and scored in the Likert tradition. *Educational and Psychological Measurement, 38,* 665–680.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 397–479). Reading MA: Addison-Wesley.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis.* Cambridge MA: MIT Press.

Cochran, W. (1952). The X^2 test of goodness-of-fit. *Annals of Mathematical Statistics, 23,* 315–345.

Cochran, W. (1954). Some methods for strengthening the common X^2 test. *Biometrics, 10,* 256–266.

Cox, M. A. A., & Placket, R. L. (1980). Small samples in contingency tables. *Biometrika, 67,* 1–13.

Deming, W. E., & Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics, 11,* 427–444.

Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests* [Introduction to the theory of psychological tests]. Bern: Huber.

Fischer, G. H. (1995). 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.

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. (1989). Extensions of the partial credit model. *Psychometrika, 54,* 635–660.

Haberman, S. J. (1977). Log-linear models and frequency tables with small expected cell counts. *Annals of Sta-*

tistics, 5, 1148–1169.

Haberman, S. J. (1979). *Analysis of qualitative data: New developments* (Vol. 2). New York: Academic Press.

Kelderman, H. (1992). Computing maximum likelihood estimates of loglinear models from marginal sums with special attention to loglinear item response theory. *Psychometrika, 57,* 437–450.

Kelderman, H. (1995). The polytomous Rasch model within the class of generalized linear symmetry models. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 307–324). New York: Springer-Verlag.

Kelderman, H., & Steen, R. (1993). *LOGIMO 1: A program for loglinear item response theory modeling* [Computer program manual]. Groningen, The Netherlands: IEC ProGAMMA.

Koehler, K. J. (1977). *Goodness-of-fit statistics for large sparse multinomials.* Unpublished doctoral dissertation, University of Minnesota, Minneapolis, School of Statistics.

Lancaster, H. O. (1961). Significance tests in discrete distributions. *Journal of the American Statistical Association, 56,* 223–234.

Larnz, K. (1978). Small-sample comparisons of exact levels for chi-square statistics. *Journal of the American Statistical Association, 73,* 412–419.

Lehman, E. L. (1983). *The theory of point estimation.* New York: Wiley.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale NJ: Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading MA: Addison-Wesley.

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

Masters, G. N. (1987). Measurement models for ordered response categories. In R. Langeheine & J. Rost (Eds.), *Latent trait and latent class models* (pp. 11–29). New York: Plenum.

Molenaar, I. W. (1995). Some background for item response theory and the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 3–14). New York: Springer-Verlag.

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement, 14,* 59–71.

Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica, 16,* 1–32.

Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York: Wiley.

Rasch, G. (1960/80). *Probabilistic models for some intelligence and attainment tests.* (Copenhagen, Danish Institute for Educational Research). Expanded edition (1980), with foreword and afterword by B. D. Wright. Chicago: The University of Chicago Press.

Rasch, G. (1961). On the general laws and meaning of measurement in psychology. In J. Neyman (Ed.), *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 5; pp. 321–333). Berkeley: University of California Press.

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy, 17,* 58–94.

Raven, J., Raven, J. C., & Court, J. H. (1991). *Manual for Raven's Progressive Matrices and Vocabulary Scales (section 1): General overview.* Oxford: Oxford Psychologists Press.

Roskam, E. E., & Jansen, P. G. W. (1984). A new derivation of the Rasch model. In E. Degreef & J. Van Buggenhaut (Eds.), *Trends in mathematical psychology* (pp. 293–308). Amsterdam: Elsevier Science Publishers.

Rost, J. (1988). Measuring attitudes with a threshold model drawing on a traditional scaling concept. *Applied Psychological Measurement, 12,* 397–409.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement,* No. 17.

Sternberg, R. J. (1990). *Metaphors of mind: Conceptions of the nature of intelligence.* Cambridge: Cambridge University Press.

van den Wollenberg, A. L. (1979). *The Rasch model and time limit tests.* Unpublished doctoral thesis, Katholieke Universiteit Nijmegen.

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

van Kuyk, J. J. (1988). Verwerven van grootte-begrippen [Acquiring size concepts]. *Pedagogische Studièn, 65,* 1–10.

Vodegel Matzen, L. (1994). *Performance on Raven's Progressive Matrices: What makes the difference?* Unpublished doctoral dissertation, University of Amsterdam, Faculty of Psychology.

Author's Address

Send requests for reprints or further information to Henk Kelderman, Division of Work and Organizational Psychology, Faculty of Psychology and Pedagogics, Vrije Universiteit Amsterdam, van de Boechorststraat 1, 1018 BT Amsterdam, The Netherlands. Email: h.kelderman@psy.vu.nl.