

Chapter 7

Exponential Family Models for Continuous Responses



Norman D. Verhelst

In memory of Arie Dirkzwager

Abstract Two models for continuous responses that allow for separation of item and person parameters are explored. One is a newly developed model which can be seen as a Rasch model for continuous responses, the other is a slight generalization of a model proposed by Müller (1987). For both models it is shown that CML-estimation is possible in principle, but practically unfeasible. Estimation of the parameters using only item pairs, a form of pseudo-likelihood estimation, is proposed and detailed expressions for first and second order partial derivatives are given. A comparison of the information function between models for continuous and for discrete observations is discussed. The relation between these models and the probability measurement developed in the 1960s is addressed as well.

7.1 Introduction

In cognitive tests (achievement tests, placement tests) or aptitude tests, as well as in personality tests and attitude tests, the response variables for the items are discrete, having a very limited number of values: often only two in cognitive tests (representing incorrect/correct) or a few as with the use of Likert scales in attitude or personality tests, and described by expressions as ‘strongly disagree’, ‘disagree’, ‘neutral’, ‘agree’, ‘strongly agree’. Note that in the latter case, the categories are considered as ordered, meaning that, e.g., the choice of the category ‘strongly agree’ is a sign of a more positive attitude (motivation, etc.) than the answer ‘agree’, and ‘agree’ is

N. D. Verhelst (✉)
Eurometrics, Tiel, The Netherlands
e-mail: Norman.verhelst@gmail.com

© The Author(s) 2019
B. P. Veldkamp and C. Sluijter (eds.), *Theoretical and Practical Advances
in Computer-based Educational Measurement*, Methodology of Educational
Measurement and Assessment, https://doi.org/10.1007/978-3-030-18480-3_7

more positive than ‘neutral’, etc., or the reverse, depending of the wording of the item stem. It is usually assumed that the direction of the order can be derived from a semantic analysis of the item stem, i.e., it is not an outcome of a statistical analysis of the response data.

Measurement models to analyze these highly discrete data are widespread and well known among psychometricians. Models from the logistic family, such as the one-two- or three-parameter logistic models for binary responses (Rasch 1960; Lord and Novick 1968), the (generalized) partial credit model (Masters 1982; Muraki 1997) and the graded response model (Samejima 1974, 1997) and models, known as ‘normal ogive’, originating in biological assay (Finney 1978), but becoming popular in psychometric circles with the seminal paper by Albert (1992) on the use of sampling methods to estimate parameters.

Although certainly less popular than models for discrete data, several interesting models have been proposed and studied for continuous data, almost all in the realm of attitude scaling and personality assessment. Also, different item formats are possible to record continuous responses. One could for example present a piece of line where the endpoints are labeled (e.g., as ‘strongly disagree’ and ‘strongly agree’) and ask the respondent to put a cross at a position that best represents his or her attitude. But there are others as well. The common feature is that all answers are bounded from below and from above and that the possible answers can be regarded as a (close approximation to a) continuous response. The precise item format, although important by itself, is not considered any further in this chapter.

Samejima (1973, 1974) proposed a (family) of continuous IRT models, all derived from the graded response model (Samejima 1969, 1997) as a limiting case when the number of category responses goes to infinity. The model is complex as it assumes a latent response on the interval $(-\infty, +\infty)$ and a transformation to an interval bounded from both sides, see Bejar (1977) for an application. One of the complicating factors in Samejima’s approach is that there are two layers of latent variables: one is the construct (attitude, self-concept) to be measured and the other is a latent response which is an unbounded continuous variable, while the observable response is continuous but bounded from below and from above. To reconcile these incompatible restrictions, a transformation from an unbounded variable to a bounded one is proposed, e.g., the logit transformation. In the model developed by Mellenbergh (1994) such a transformation is not used: the continuous response variable is modeled with a one-factor model:

$$X_{ij} = \mu_j + \lambda_j \theta_i + \varepsilon_{ij},$$

where X_{ij} is the continuous observed response of respondent i to item j , μ_j is the easiness parameter of the item, λ_j the discrimination parameter, θ_i the value of the latent variable for respondent i , and ε_{ij} the continuous residual, which is usually but not necessarily normally distributed with mean zero and variance σ^2 . A comparison between the models, of Samejima and Mellenbergh, can be found in Ferrando (2002).

More recently, an interesting family of models has been proposed where the observed response variable, rescaled to the unit interval, follows a beta distribution, and can therefore be used to model monotone as well as single peaked densities (Noel and Dauvier 2007; Noel 2014). The work of Noel (2017) concentrated on unfolding models for personality traits, emotions and behavioral change.

Neither of the above models, however, are exponential family models, or do allow for separation of person and item parameters. The only published model for continuous responses which is an exponential family model and allows for parameter separation is published by Müller (1987). This model and a newly proposed one are the main topics of this chapter.

Although all the models discussed so far consider the use of continuous response variables, none of them is used for cognitive tests, like achievement tests or placement tests. For these tests considerable attention has been paid to continuous responses, especially for multiple choice tests, not in an IRT framework, but in what was called probability measurement. In this approach, the respondent does not have to pick a single alternative, but has to express his or her degree of belief or subjective probability that each of the response alternatives is the right answer (De Finetti 1965, 1970; Van Naerssen 1961; Toda 1963; Roby 1965). The attention in the sixties was directed to the problem of a good or 'admissible' scoring function, i.e., a scoring rule such that the continuous responses will reflect the true subjective probabilities. See Rippey (1970) for a comparative study of different scoring rules.

An important contribution in this area was made by Shuford et al. (1966) who showed that there was only one scoring rule (i) which maximized the expected score if and only if the continuous answers were equal to the subjective probabilities and (ii) where the score only depends on the response for the correct alternative and not on the distribution across the incorrect ones. This scoring rule is the logarithmic one, and of course has an evident disadvantage: if the continuous response to the correct alternative is zero, the score is minus infinity and can never be repaired by any finite number of (partially) correct responses. Shuford et al. were aware of this anomaly and proposed a slight modification of the logarithmic rule by a truncation on the observable variable: responses at or below 0.01 got a fixed penalty of -2 . Dirkwager (1997, 2001)¹ provided an elegant way to avoid very large penalties and at the same time to have a good approximation to the original logarithmic scoring rule. The approximation is dependent on a tuning parameter. Details on the scoring function and a quite extensive application of the system developed by Dirkwager can be found in Holmes (2002).

Probably the main factor that hampered the application of these ideas was the poor development of suitable computer systems. Contrary to the situation with scales for personality traits, emotions and attitudes, where a continuous answer can be elicited, for example, by asking to put a mark on a line where the two end points are indicated by labels such as 'strongly disagree' and 'strongly agree' and where the mark expresses best the respondents position, the continuous answers for multiple choice

¹Many of the writings of the late Arie Dirkwager are not easy to find. Holmes (2002) gives quite detailed reports of his writings.

questions are multivariate. Nowadays modern computers are widespread and constructing interfaces for administering multiple choice tests with continuous answers is hardly a serious challenge.

In Sect. 7.2 a new model, a simple continuous Rasch model is introduced and parameter estimation is discussed. Section 7.3 treats a slight generalization of Müller's model and gives the details of parameter estimation along the lines sketched, but not elaborated in his 1987 article. Section 7.4 handles the problem of comparisons of information functions for different models and in Sect. 7.5 a discussion is started about the relation between IRT models for continuous responses in multiple choice tests and the scoring functions which were studied in the sixties.

7.2 A Rasch Model for Continuous Responses

7.2.1 The Model

Much in the spirit of the probability measurement approaches, let us imagine that a test consisting of k multiple choice items has been answered in a continuous way by assigning a non-zero number to each alternative under the restriction that the sum of these numbers across alternatives equals one. At this moment we leave the precise instructions to the respondents a bit vague, except for the fact that they know that the higher the number assigned to the right answer, the higher their score will be, i.e., there is a monotone increasing relationship between answer and the score. More on this will be said in the discussion section.

In the item response function of the Rasch model for binary data, the denominator $1 + \exp(\theta - \beta_i)$ has the role of a normalizing constant, i.e., the sum of all possible numerators, and guarantees that the sum of the probabilities of all possible answers equals one. If one considers a similar model for continuous responses in the (closed) unit interval, one arrives readily at the conditional probability density function

$$f_i(r_i|\theta) = \frac{\exp[r_i(\theta - \eta_i)]}{\int_0^1 \exp[t(\theta - \eta_i)] dt}, \quad (r_i \in [0, 1]), \quad (7.1)$$

where r_i is a realization of the random variable R_i , the answer given to the correct alternative of item i , θ is the underlying ability, continuous and unbounded and η_i is an item parameter, representing the difficulty as in the Rasch model for binary data².

The denominator of (7.1) has a closed form solution. If $\theta = \eta_i$, the integral clearly equals one, as well as the numerator of (7.1). Using α_i as a shorthand for $\theta - \eta_i$, and the result from calculus that

²Another symbol for the difficulty parameter is used to avoid the suggestion that the difficulty parameters in the binary and continuous mode should be equal.

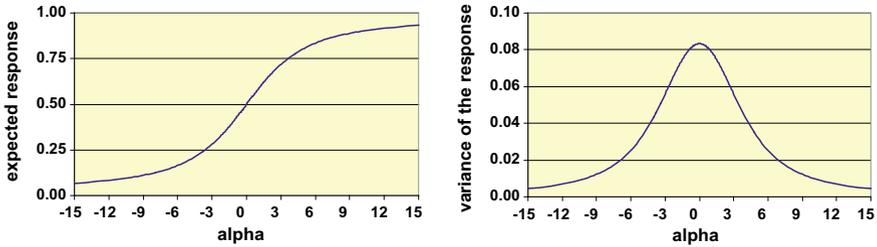


Fig. 7.1 Expected value (left) and variance (right) of R_i as a function of $\alpha_i = \theta - \eta_i$ in the Rasch model for continuous responses

$$\int \exp(t\alpha) dt = \frac{1}{\alpha} \exp(t\alpha),$$

(7.1) can also be written as

$$f_i(r_i|\theta) = \begin{cases} \frac{(\theta - \eta_i) \exp[r_i(\theta - \eta_i)]}{\exp[(\theta - \eta_i)] - 1} & \text{if } \theta \neq \eta_i \text{ or } \alpha_i \neq 0, \\ 1 & \text{if } \theta = \eta_i \text{ or } \alpha_i = 0. \end{cases} \tag{7.2}$$

The regression of the response variable R_i on θ is given by

$$E(R_i|\theta) = \frac{\int_0^1 t \exp(t\alpha_i) dt}{\int_0^1 \exp(t\alpha_i) dt} = \begin{cases} \frac{e^{\alpha_i}(\alpha_i - 1) + 1}{\alpha_i(e^{\alpha_i} - 1)} & \text{if } \alpha_i \neq 0, \\ 0.5 & \text{if } \alpha_i = 0, \end{cases} \tag{7.3}$$

where α_i is used as a shorthand for $\theta - \eta_i$. (Note that for the case $\alpha_i = 0$, the two integrals in (7.3) have a trivial solution.) In Fig. 7.1 (left panel) a graphical representation of the regression function is given. Notice that the horizontal axis is the difference between θ and the item parameter, and therefore the regression graph itself (with θ on the horizontal axis) will have the same form as the curve in Fig. 7.1, and for different items the curves will be shifted horizontally with respect to each other, just as the trace lines in the Rasch model for binary items. Without giving formal proofs we state some characteristics of the regression function:

1. $\lim_{\theta \rightarrow -\infty} E(R_i|\theta) = 0$,
2. $\lim_{\theta \rightarrow +\infty} E(R_i|\theta) = 1$,
3. $E(R_i|\theta)$ is monotonically increasing in θ .

One might be wondering about the numbers along the horizontal axis. In the model for binary data the graph of the item response function (which is a regression function) is usually very close to zero or to one if $|\theta - \beta_i| > 3$, while we see here a substantial difference between the expected value of the response and its lower or higher asymptote for $|\alpha|$ as large as 15. In the section about the information function we will come back to this phenomenon and give a detailed account of it.

It is easy to see from (7.1) or (7.2) that it is an exponential family density with r_i as sufficient statistic for θ . In exponential families the (Fisher) information is the variance of the sufficient statistic, and the expression for this variance is

$$\text{Var}(R_i|\theta) = \begin{cases} \frac{1}{\alpha_i^2} - \frac{e^{\alpha_i}}{(e^{\alpha_i} - 1)^2} & \text{if } \alpha_i \neq 0, \\ \frac{1}{12} & \text{if } \alpha_i = 0, \end{cases} \tag{7.4}$$

where the value 1/12 is the limit of the expression above it in (7.4), or the variance of a uniformly distributed variable in the unit interval. In Fig. 7.1 (right panel) a graph of the variance function is displayed.

7.2.2 Parameter Estimation

To make the model complete, we have to add an assumption about the dependence structure between item responses. As is common in most IRT models, we will assume that the responses to the items are conditionally (or locally) independent. With this assumption the model is still an exponential family and the sufficient statistic for the latent variable θ is $R = \sum_i R_i$. Therefore conditional maximum likelihood (CML) estimation is in principle possible and in this section the practical feasibility of CML is explored.

The conditional likelihood³ of the data, as a function of the parameter vector $\eta = (\eta_1, \dots, \eta_k)$ given the value of the sufficient statistic for θ is given by

$$L(\eta|R = r) = \frac{\exp[-\sum r_i \eta_i]}{\int_A \prod_i \exp(-t_i \eta_i) dt_i} \tag{7.5}$$

where \int_A denotes the multiple integral over all score vectors (t_1, \dots, t_k) such that $\sum_i t_i = r$. As a general symbol for this multiple integral, we will use $\gamma(r; k)$, where the first argument refers to the score r and the second argument denotes the number of items in the test.

To see the complexity of these γ -functions, consider first the case where $k = 2$. If $r \leq 1$, the score on any of the two items cannot be larger than r , but each score can be zero. So, let t_1 be the score on the first item; then t_1 can run from 0 to r , and of course $t_2 = r - t_1$ is linearly dependent on t_1 . If $r > 1$, each score is bounded not to be smaller than $r - 1$, and of course each score cannot be larger than 1. Taking these considerations into account it is easily verified that

$$\gamma(r, 2) = \int_{\max(r-1, 0)}^{\min(r, 1)} \exp[-(r - t_1)\eta_2] \exp(-t_1 \eta_1) dt_1 \tag{7.6}$$

³For continuous data, the likelihood is proportional to the density of the observed data.

Using similar considerations as in the case with $k = 2$, one can verify that

$$\gamma(r, 3) = \int_{\max(r-2, 0)}^{\min(r, 1)} \exp(-t_2 \eta_2) \int_{\max(r-t_2-1, 0)}^{\min(r-t_2, 1)} \exp[(r-t_2-t_1)\eta_3] \exp(-t_1 \eta_1) dt_1 dt_2,$$

and in general we can write

$$\gamma(r, k) = \int_{A_{k-1}}^{B_{k-1}} f_{k-1}(t_{k-1}) \dots \int_{A_i}^{B_i} f_1(t_1) \dots g(t_1, \dots, t_{k-1}) dt_1 \dots dt_{k-1}, \quad (7.7)$$

where

$$\begin{aligned} A_i &= \max\left(0, r - (k-i) - \sum_{j=1}^{i-1} t_{k-j}\right), \\ B_i &= \min\left(1, r - \sum_{j=1}^{i-1} t_{k-j}\right), \\ f_i(t_i) &= \exp(-t_i \eta_i), \\ g(t_1, \dots, t_{k-1}) &= \exp\left[-\eta_k \left(r - \sum_{j=1}^{k-1} t_j\right)\right]. \end{aligned}$$

It will be clear that evaluation of (7.7), although an explicit solution exists, is very unpractical, since in all integrals a distinction is to be made between two possible minima and two possible maxima in every integration. To illustrate this, consider the solution of (7.6), assuming that $\eta_1 \neq \eta_2$:

$$\gamma(r, 2) = \begin{cases} \frac{1}{\eta_2 - \eta_1} [\exp(-r \eta_1) - \exp(-r \eta_2)], & (0 < r < 1), \\ \frac{1}{\eta_2 - \eta_1} [\exp(-\eta_1 - (r-1)\eta_2) - \exp(-\eta_2 - (r-1)\eta_1)], & (1 \leq r < 2). \end{cases} \quad (7.8)$$

If $\gamma(r, k)$ is evaluated, k different expressions will result, which are too complicated to work with. Therefore, the maximization of the conditional likelihood function is abandoned; instead recourse is taken to a pseudo-likelihood method, where the product of the conditional likelihood of all pairs of variables is maximized (Arnold and Strauss 1988, 1991; Cox and Reid 2004). This means that the function

$$PL(\eta) = \prod_{i < j} \frac{\exp(-r_i \eta_i - r_j \eta_j)}{\gamma_{r_{ij}}(\eta_i, \eta_j)}$$

will be maximized. The variable r_{ij} is defined by

$$r_{ij} = r_i + r_j, \quad (i \neq j) \quad (7.9)$$

and $\gamma_{r_{ij}}(\eta_i, \eta_j)$ is the explicit notation of $\gamma(r_{ij}, 2)$ with η_i and η_j as arguments.

At this point, it proves useful to reparametrize the model. Define

$$\eta_{ij} = \frac{\eta_i + \eta_j}{2} \quad (7.10)$$

and

$$\varepsilon_{ij} = \eta_{ij} - \eta_i = \frac{\eta_j - \eta_i}{2}. \quad (7.11)$$

It follows immediately that $\varepsilon_{ij} = -\varepsilon_{ji}$. Using definitions (7.10) and (7.11) and Eq. (7.8), the factor of the PL-function referring to the item pair (i, j) can be written as

$$PL(\varepsilon_{ij}) = \begin{cases} \frac{\varepsilon_{ij} \exp[(r_i - r_j)\varepsilon_{ij}]}{\sinh(r_{ij}\varepsilon_{ij})}, & (0 < r_{ij} < 1), \\ \frac{\varepsilon_{ij} \exp[(r_i - r_j)\varepsilon_{ij}]}{\sinh[(2 - r_{ij})\varepsilon_{ij}]}, & (1 \leq r_{ij} < 2), \end{cases} \quad (7.12)$$

if $\varepsilon_{ij} \neq 0$, this is, if $\eta_i \neq \eta_j$. If j and i are interchanged (7.12) does not change because $\varepsilon_{ij} = -\varepsilon_{ji}$ and $\sinh(-x) = -\sinh(x)$. If $\eta_i = \eta_j$, the solution can be found directly from (7.5). It is given by

$$PL(0) = \lim_{\varepsilon_{ij} \rightarrow 0} PL(\varepsilon_{ij}) = \begin{cases} \frac{1}{r_{ij}}, & (0 < r_{ij} < 1), \\ \frac{1}{2 - r_{ij}}, & (1 \leq r_{ij} < 2). \end{cases} \quad (7.13)$$

In the sequel, only reference will be made to (7.12), but (7.13) has to be used if appropriate.

Although the events $(r_{ij} = 0)$ and $(r_{ij} = 2)$ both have probability zero, they can occur in a data set. The joint conditional density of (r_i, r_j) given $(r_{ij} = 0)$ or $(r_{ij} = 2)$, however, is independent of the item parameters, and can therefore be eliminated from the data set; only values in the open interval $(0, 2)$ are to be considered. This is similar to the fact that in the Rasch model for binary data response patterns with all zeros or all ones can be removed from the data without affecting the CML-estimates of the item parameters.

To get rid of the double expression in the right-hand side of (7.12) define the indicator variable A_{ij} , with realizations a_{ij} , as

$$A_{ij} = \begin{cases} 0 & \text{if } 0 \leq r_{ij} < 1, \\ 1 & \text{if } 1 \leq r_{ij} \leq 2, \end{cases}$$

and define the random variable B_{ij} , with realizations b_{ij} , as

$$B_{ij} = R_{ij} + 2A_{ij}(1 - R_{ij}) \quad (7.14)$$

Using (7.14), (7.12) can be rewritten as

$$PL(\varepsilon_{ij}) = \frac{\varepsilon_{ij} \exp[(r_i - r_j)\varepsilon_{ij}]}{\sinh(b_{ij}\varepsilon_{ij})} \quad (7.15)$$

and (7.13) can be rewritten as

$$PL(0) = \lim_{\varepsilon_{ij} \rightarrow 0} PL(\varepsilon_{ij}) = \frac{1}{b_{ij}}$$

Taking the logarithm of (7.15) and differentiating with respect to ε_{ij} yields

$$\frac{d \ln[PL(\varepsilon_{ij})]}{d\varepsilon_{ij}} = (r_i - r_j) + \frac{1}{\varepsilon_{ij}} - \frac{b_{ij}}{\tanh(b_{ij}\varepsilon_{ij})}, \quad (7.16)$$

and differentiating a second time gives

$$\frac{d^2 \ln[PL(\varepsilon_{ij})]}{d\varepsilon_{ij}^2} = -\frac{1}{\varepsilon_{ij}^2} + \frac{b_{ij}^2}{\sinh^2(b_{ij}\varepsilon_{ij})}. \quad (7.17)$$

If $\varepsilon_{ij} = 0$, (7.16) and (7.17) are undefined, but can be replaced by their limits:

$$\lim_{\varepsilon_{ij} \rightarrow 0} \frac{d \ln[PL(\varepsilon_{ij})]}{d\varepsilon_{ij}} = r_i - r_j$$

and

$$\lim_{\varepsilon_{ij} \rightarrow 0} \frac{d^2 \ln[PL(\varepsilon_{ij})]}{d\varepsilon_{ij}^2} = -\frac{b_{ij}^2}{3}.$$

In order to obtain estimates of the original η -parameters, the restrictions, defined by (7.11) have to be taken in account. Define the $k(k-1)/2 \times k$ matrix K by

$$K(ij, \ell) = \begin{cases} -1/2 & \text{if } \ell = i, \\ 1/2 & \text{if } \ell = j, \\ 0 & \text{Otherwise} \end{cases} \quad (7.18)$$

where the subscript ij of the rows refers to the item pair (i, j) . Define $\varepsilon = (\varepsilon_{12}, \varepsilon_{13}, \dots, \varepsilon_{ij}, \dots, \varepsilon_{k-1, k})$ and $\eta = (\eta_1, \dots, \eta_k)$, then it follows immediately from (7.11) to (7.18) that

$$\varepsilon = K\eta. \quad (7.19)$$

It is immediately clear that

$$\frac{\partial \ln PL(\eta)}{\partial \eta} = K' \frac{\partial \ln PL(\varepsilon)}{\partial \varepsilon}, \quad (7.20)$$

and

$$\frac{\partial^2 \ln PL(\eta)}{\partial \eta \partial \eta'} = K' \frac{\partial^2 \ln PL(\varepsilon)}{\partial \varepsilon \partial \varepsilon'} K, \quad (7.21)$$

where the elements of the partial derivatives with respect to ε are given by (7.16). The matrix of second partial derivatives in the right-hand side of (7.21) is a diagonal matrix, whose diagonal elements are defined by (7.17). For a sample of n response patterns, the gradient and the matrix of second order partial derivatives of the PL-function is simply the sum over response patterns of expressions given by the right-hand members of Eqs. (7.20) and (7.21), respectively. The model, however, is not identified unless a normalization restriction is imposed on the η -parameters, e.g., $\eta_k = 0$. This amounts to dropping the last element of the gradient vector and the last rows and columns from the matrix of second order partial derivatives.

Initial estimates can be found by equating the right-hand member of (7.16) to zero, and solving as a univariate problem, i.e., ignoring the restrictions (7.19). Applying (7.19) then yields least squares estimates of η , which can be used as initial estimates of the item parameters.

Standard errors can be found by the so-called sandwich method. Define \mathbf{g}_v as the vector of first partial derivatives of the pseudo-likelihood function for respondent v and H as the matrix of second partial derivatives (for all respondents jointly). All vectors \mathbf{g}_v and the matrix H are evaluated at the value of the pseudo-likelihood estimates of the parameters. Then, the asymptotic variance-covariance matrix can be estimated by (Cox and Read 2004, p. 733)

$$H^{-1} \left[\sum_v \mathbf{g}_v \mathbf{g}_v' \right] H^{-1}.$$

7.3 An Extension of the Müller Model

7.3.1 The Model

As Samejima derived her family of continuous models as limiting cases of the graded response model when the number of possible graded responses goes to infinity, Müller considers a limiting case of Andrich's (1982) rating scale model when the number of thresholds tends to infinity. The category response function of the rating scale model (with possible answers $0, 1, \dots, m$) is given by

$$P(X_i = j|\theta^*) = \frac{\exp[j\alpha_i^* + j(m - j)\delta^*]}{\sum_{h=0}^m \exp[h\alpha_i^* + h(m - h)\delta^*]} \tag{7.22}$$

where $\alpha_i^* = \theta^* - \eta_i^*$, the difference between the latent value and an item specific location parameter, while δ^* is half the (constant) difference between any two consecutive thresholds^{4,5} If the answer R_i (with realizations r_i) is elicited by asking the respondent to put a mark on a piece of line with length $d(>0)$ and midpoint c and where the two endpoints are labeled, then the response can be modeled by the density

$$f_i^*(r_i|\theta) = \frac{\exp[r_i\alpha_i + r_i(2c - r_i)\delta]}{\int_{c-d/2}^{c+d/2} \exp[t\alpha_i + t(2c - t)\delta] dt}, \quad \left(r_i \in \left[c - \frac{d}{2}, c + \frac{d}{2} \right] \right), \tag{7.23}$$

with $\alpha_i = \theta - \eta_i$. Moreover, Müller shows that the thresholds are uniformly distributed in the interval $[\eta_i - \delta d, \eta_i + \delta d]$.

Of course, the length of the line in the response format can be expressed in arbitrary units and with an arbitrary reference point, so that we can assume without loss of generality that $c = 0.5$ and $d = 1$. And as a slight extension of Müller’s model, it will be assumed that the δ -parameter can vary across items. This gives the density equation we will be using in the present section:

$$f_i(r_i|\theta) = \frac{\exp[r_i\alpha_i + r_i(1 - r_i)\delta_i]}{\int_0^1 \exp[t\alpha_i + t(1 - t)\delta_i] dt}, \tag{7.24}$$

and by completing the square, the numerator of (7.24) can be written as

$$\exp[r_i\alpha_i + r_i(1 - r_i)\delta_i] = \exp\left[-\delta_i\left(r_i - \frac{\alpha_i + \delta_i}{2\delta_i}\right)^2\right] \times \exp\left[\frac{(\alpha_i + \delta_i)^2}{4\delta_i}\right]. \tag{7.25}$$

The second factor in the right-hand side of (7.25) is independent of r_i and will cancel when we substitute the right-hand side of (7.25) in (7.24). Defining

$$\mu_i(\theta) = \frac{\alpha_i + \delta_i}{2\delta_i} \text{ and } \sigma_i^2 = \frac{1}{2\delta_i},$$

(7.24) can be written as

⁴The “*” in Eq. (7.22) is introduced to avoid the suggestion that parameters and variables in the discrete and the continuous model are identical.

⁵In the derivation of the rating scale model, Andrich assumes that a response in category j means that the j most left positioned thresholds out of m have been passed and the $(m - j)$ rightmost ones not. The location parameter η_j is the midpoint of the m thresholds.

$$f_i(r_i|\theta) = \frac{\sigma_i^{-1} \varphi\left[\frac{r-\mu_i(\theta)}{\sigma_i}\right]}{\Phi\left[\frac{1-\mu_i(\theta)}{\sigma_i}\right] - \Phi\left[\frac{-\mu_i(\theta)}{\sigma_i}\right]} \tag{7.26}$$

with $\varphi(\cdot)$ and $\Phi(\cdot)$ denoting the standard normal density and probability functions, respectively. One easily recognizes (7.26) as the probability density function of the truncated normal distribution (Johnson and Kotz 1970). The regression of the response on the latent variable and the variance of the response are given next. Using D_i as shorthand for the denominator of (7.26), i.e.,

$$D_i = \Phi\left[\frac{1-\mu_i(\theta)}{\sigma_i}\right] - \Phi\left[\frac{-\mu_i(\theta)}{\sigma_i}\right],$$

and

$$z_{0i} = \frac{-\mu_i(\theta)}{\sigma_i} \text{ and } z_{1i} = \frac{1-\mu_i(\theta)}{\sigma_i},$$

the regression function is given by

$$E(R_i|\theta) = \mu_i(\theta) + \frac{\varphi(z_{0i}) - \varphi(z_{1i})}{D_i} \sigma_i,$$

and the variance by

$$Var(R_i|\theta) = \sigma_i^2 \left[1 + \frac{z_{0i}\varphi(z_{0i}) - z_{1i}\varphi(z_{1i})}{D_i} - \left(\frac{\varphi(z_{0i}) - \varphi(z_{1i})}{D_i} \right)^2 \right].$$

The regression function is displayed in the left panel of Fig. 7.2 for three different values of the δ -parameter. Some comments are in order here:

1. The computation of the regression and variance is tricky for extreme values of z_{0i} and z_{1i} . Using the standard implementation of the normal probability function

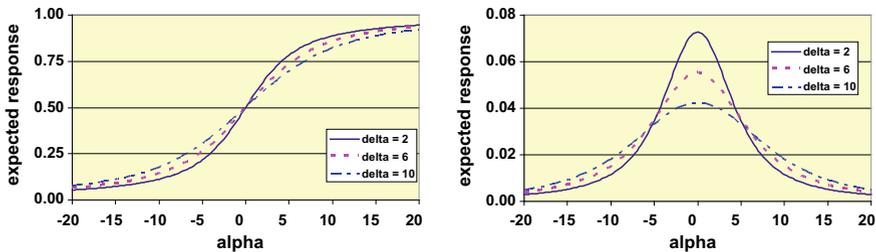


Fig. 7.2 Expected value (left) and variance (right) of R_i as a function of $\alpha_i = \theta - \eta_i$ in Müller’s model

in EXCEL or R generates gross errors. Excellent approximations are given in Feller (1957, p. 193).

2. Three values of the δ -parameter have been used, resulting in flatter curves for the higher values. Therefore, the parameter δ can be interpreted as a discrimination parameter, lower values yielding higher discrimination. A discussion on the model properties when $\delta \rightarrow 0$ can be found in Müller (1987).
3. Just as in the Rasch model for continuous responses, the numbers along the horizontal axis are quite different from the ones usually displayed for the item response functions of the common Rasch model. Further comments will be given in Sect. 7.4.

The right-hand panel of Fig. 7.2 displays the variance of the response function for the same three values of δ and for the same values of α . The figures are symmetric around the vertical axis at $\alpha = 0$. As this model is an exponential family, the variance of the response is also the Fisher information.

To gain more insight in the information function, one can use a reparameterization of the rating scale model, where the scores are brought back to the unit interval, i.e., by dividing the original integer valued scores by m_i , the original maximum score. Müller shows that this results in the discrete model with category response function⁶

$$f_i^*(r_i|\theta) = \frac{\exp[r_i\alpha_i + r_i(1 - r_i)\delta_i]}{\sum_{j=0}^m \exp[r_j\alpha_i + r_j(1 - r_j)\delta_i]}, \quad \left(r_i = 0, \frac{1}{m}, \frac{2}{m}, \dots, 1 \right). \quad (7.27)$$

It is easy to see that (7.24) is the limiting case of (7.27) as $m \rightarrow \infty$. In Fig. 7.3 the information functions are plotted for some finite values of m and for the limiting case, labeled as ‘continuous’. In the left-hand panel, the discrimination parameter δ_i equals 2 (high discrimination), in the right-hand panel it equals 10 (low discrimination); the value of η_i is zero in both cases such that $\alpha_i = \theta$.

The collection of curves can be described briefly as follows:

1. the maxima of the curves are lower with increasing values of m ;
2. the tails are thicker with increasing values of m ;

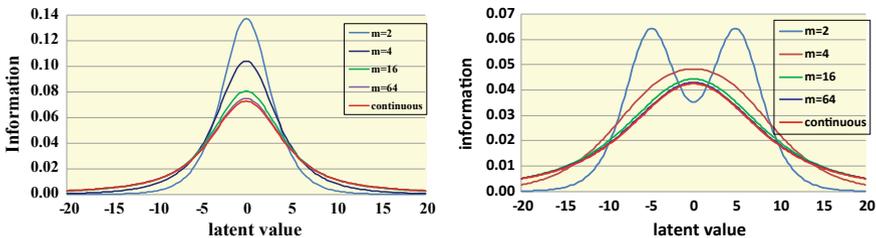


Fig. 7.3 Information functions for different values of m (left: $\delta_i = 2$; right: $\delta_i = 10$)

⁶In Müller’s article the admitted values for r_i are slightly different, but the difference is not important.

3. for low values of m the curves are not necessarily unimodal: see the case for $m = 2$ en $\delta_i = 10$;
4. for $m = 64$ and $m \rightarrow \infty$, the curves are barely distinguishable.

One should be careful with conclusions drawn from the graphs in Fig. 7.3: it does not follow, for example, from this figure that, if one switches from a discrete format of an item to a continuous one, that the discrimination or the location parameter will remain invariant. An example will be given in Sect. 7.4.

7.3.2 Parameter Estimation

Müller (1987) proposes the parameter estimation using only pairs of items, but does not give details. In the present section the method is explained; the technical expressions to compute the gradient and the matrix of second partial derivatives are given in the Appendix to this chapter.

As with the continuous Rasch model, definition (7.9), $r_{ij} = r_i + r_j$ is used and the reparameterization

$$\varepsilon_i = \delta_i - \eta_i.$$

is used. The problems with obtaining CML estimates of the parameters are the same as with the Rasch model for continuous responses (and augmented with numerical problems in evaluating the probability function of the truncated normal distribution for extreme values of the argument). Therefore the pseudo-likelihood function, considering all pairs of items, is studied here.

The conditional likelihood function for one pair of items is given by

$$f_{ij}(r_i, r_j | r_{ij}) = \frac{\exp(r_i \varepsilon_i + r_j \varepsilon_j - r_i^2 \delta_i - r_j^2 \delta_j)}{\int_{m_{ij}}^{M_{ij}} \exp[t \varepsilon_i + (r_{ij} - t) \varepsilon_j - t^2 \delta_i - (r_{ij} - t)^2 \delta_j] dt} \quad (7.28)$$

where the bounds of the integral in the denominator depend on the value of r_{ij} :

$$m_{ij} = \max(r_{ij} - 1, 0),$$

$$M_{ij} = \min(r_{ij}, 1).$$

Along the same lines of reasoning as followed in Sect. 3.1, (7.28) can also be shown to be a density of the truncated normal distribution, i.e.,

$$f_{ij}(r_i, r_j | r_{ij}) = \frac{\exp\left[-\frac{(r_i - \mu_{ij})^2}{2\sigma_{ij}^2}\right]}{\int_{m_{ij}}^{M_{ij}} \exp\left[-\frac{(t - \mu_{ij})^2}{2\sigma_{ij}^2}\right] dt} = \frac{\sigma_{ij}^{-1} \varphi(x_{ij})}{\Phi(Z_{ij}) - \Phi(z_{ij})}, \quad (7.29)$$

where

$$\begin{aligned}\sigma_{ij}^2 &= \frac{1}{2(\delta_i + \delta_j)}, \\ \mu_{ij} &= \sigma_{ij}^2(\varepsilon_i - \varepsilon_j + 2r_{ij}\delta_j), \\ x_{ij} &= (r_i - \mu_{ij})/\sigma_{ij}, \\ Z_{ij} &= (M_{ij} - \mu_{ij})/\sigma_{ij}, \\ z_{ij} &= (m_{ij} - \mu_{ij})/\sigma_{ij}.\end{aligned}$$

Although the function μ_{ij} is not symmetric in i and j , it follows readily that $\mu_{ij} + \mu_{ji} = r_{ij} + r_{ji}$. Using this, it is not hard to show that the conditional response density (7.29) does not change if i and j are interchanged.

The function to be maximized is the logarithm of the pseudo-likelihood function:

$$\sum_v \ln PL_v[\varepsilon, \delta; (r_{v1}, \dots, r_{vk})] = \sum_v \sum_{i < j} \ln f_{ij}(r_{vi}, r_{vj} | r_{vij}), \quad (7.30)$$

where ε and δ denote the vectors of ε - and δ -parameters, r_{vi} is the response of respondent v to item i and k is the total number of items. Deriving the expressions for the first and second partial derivatives of (7.30) is not difficult, but rather tedious, the expressions are given in the Appendix to this chapter.

7.4 Comparison of Information Functions Across Models

7.4.1 The Unit of the Latent Variable

Müller (1987, p. 173) compares the information functions for the rating scale model and his own model for continuous data, assuming that the latent variable underlying the two models is the same, but this assumption needs not to be correct for two quite different reasons.

1. The first and most important threat to a meaningful comparison is the silent assumption that the same trait is measured because the items are the same; only the item format is different. For example, in a personality inventory, in the discrete case the respondent may be asked to express the applicability of a statement to him/herself on a discrete 5-point scale where each possible answer is labeled by descriptions as ‘certainly not applicable’, ‘not applicable’, ‘neutral’ (whatever this may mean), ‘applicable’ and ‘certainly applicable’, while in the continuous case the respondent is asked to express the applicability of the statement by putting a mark on a piece of line where only the two end points are labeled, as (e.g.,) ‘not applicable at all’ and ‘strongly applicable’. In the discrete case tendencies to the middle may strongly influence to prefer the middle category,

while this tendency may be a less important determinant in the continuous case. In principle there is only one way to have good answers on this important question of construct validity: one has to collect data under both item formats, and estimate the correlation (disattenuated for unreliability) between the traits measured using either of the two formats.

2. But even if one can show that the traits measured using different item formats are the same (i.e., their latent correlation equals one), there is still the problem of the origin and the unit of the scale. It is easily understood that the origin of the scale is arbitrary, as most of the IRT models define their item response functions using a difference between the trait value and a location parameter, such that adding an arbitrary constant to both does not change response probabilities (or densities as is easily seen from (7.1)). But understanding that the unit of the scale is arbitrary as well is not so easy. A historic example of this is the claim in Fischer (1974) that the scale identified by the Rasch model is a ‘difference scale’, a scale where the origin is arbitrary but the unit fixed, while the same author (Fischer 1995) came, in a complicated chapter, to the conclusion that the scale measured by the Rasch model is an interval scale, with arbitrary origin and unit. In the common Rasch model the unit is chosen by fixing the discrimination parameter (common to all items) at the value of 1, but any other positive value may be chosen. Suppose that in the discrete model, one chooses $c \neq 1$, then one can replace $\theta - \beta_i$ with $c(\theta^* - \beta_i^*)$ where $\theta^* = \theta/c$ and $\beta_i^* = \beta_i/c$ and c the common (arbitrarily chosen) discrimination value.

With the continuous models, however, the choice of the unit of measurement (of the latent variable) is also influenced by the bounds of the integral. We illustrate this with the Rasch model for continuous responses. Suppose data are collected by instructing the respondents to distribute M tokens (with M large enough such that the answer can safely be considered as continuous) among the alternatives of a multiple choice question and let U_i (with realizations u_i) be the number of tokens assigned to the correct alternative of item i . Then, the model that is equivalent to (7.1) is given by

$$f_i(u|\theta^*) = \frac{\exp[u(\theta^* - \eta_i^*)]}{\int_0^M \exp[y(\theta^* - \eta_i^*)] dy}, \quad (u \in [0, M]).$$

If we want to rescale the response variable by changing it to a relative value, i.e., $R_i = U_i/M$, then we find

$$\begin{aligned} f_i(r_i|\theta^*) &= \frac{\exp[r_i M(\theta^* - \eta_i^*)]}{M \int_0^1 \exp[t M(\theta^* - \eta_i^*)] dt} \\ &= \frac{\exp[r_i(\theta - \eta_i)]}{M \int_0^1 \exp[t(\theta - \eta_i)] dt} \\ &= \frac{1}{M} f_i(y_i|\theta), \quad (y_i = r_i/M \in [0, 1]) \end{aligned}$$

with $\theta = M\theta^*$ and $\eta_i = M\eta_i^*$. So we see that changing the bounds of the integration in (7.1) changes the density by a constant and at the same time the unit of the underlying scale. Changing the density is not important as it will only affect the likelihood by a constant, but at the same time the unit of the underlying scale is changed. Choosing zero and one as the integration bounds in (7.1) is an arbitrary choice, and therefore the unit of measurement is arbitrary as well.

Exactly the same reasoning holds for Müller's model where in Eq. (7.23) the constants c and d (>0) are arbitrary, but will influence the unit of measurement of the latent variable.

7.4.2 An Example

A simple example to see how the information function depends on the unit of measurement is provided by the Rasch model for binary data. Assuming, as in the previous section, that the common discrimination parameter is indicated by c , the information function for a single item is given by

$$I_i(\theta) = c^2 f_i(\theta)[1 - f_i(\theta)],$$

meaning that doubling c will quadruple the information, so that this gives the impression that the information measure is arbitrary. But one should keep in mind that doubling the value of c will at the same time halve the standard deviation (SD) of the distribution of θ or divide its variance by a factor four. This means that the information measure can only have meaning when compared to the variance of the θ -distribution. So, if we want to compare information functions across models, we must make sure that the latent variables measured by the models are congeneric, i.e., their correlation must be one, but at the same time they must be identical, i.e., having the same mean and variance. This is not easy to show empirically, but we can have an acceptable approximation as will be explained by the following example.

At the department of Psychology of the University of Amsterdam, all freshmen participate (compulsorily) in the so-called test week⁷: during one week they fill in a number of tests and questionnaires and take part as subjects in experiments run at the department. One of the questionnaires presented is the Adjective Check List (ACL, Gough & Heilbrunn, 1983; translated into Dutch by Hendriks et al. 1985), where a number of adjectives (363) is presented to the test takers. The task for the student is to indicate the degree to which each adjective applies to him/herself. The standard administration of the test asks the students to indicate the degree of applicability on a five point scale. In 1990, however, two different test forms were administered, each to about half of the students. In the first format, only a binary response was asked for (not applicable/applicable); in the second format, the student was asked to mark the degree of applicability on a line of 5.7 cm, with the left end corresponding to 'not

⁷At least, this was the case until 1990, the year where the continuous data were collected.

applicable at all' and the right end to 'completely applicable'. The score obtained by a student is the number of millimeters (m) of the mark from the left end of the line segment. This number m was transformed into a response (in model terms) by the transformation

$$r = \frac{m + 0.5}{58}.$$

Seven of these adjectives, fitting in the framework of the Big Five, were taken as relevant for the trait 'Extraversion' and these seven are used in the present example. For the polytomous data, the responses collected in the years 1991 through 1993 were pooled, yielding 1064 records. For the continuous data 237 records were available. Dichotomous data are not used here. All these records are completely filled in; some 2% of the original records with one or more omissions have been removed. The polytomous data have been analyzed with the Partial Credit Model (PCM) using CML estimation for the category parameters for ordered categories running from 0 to 4; the continuous data were analyzed with the extension of Müller's model as discussed in Sect. 7.3, using the maximum pseudo-likelihood method. After the estimation of the item parameters, the mean and SD of the θ -distribution were estimated, while keeping the item parameters constant at their previous estimates. For both models a normal distribution of the latent variable was assumed.

Parameter estimates are displayed in Table 7.1. The adjectives were presented to the students in Dutch; the English translation is only put there as information and did not have any influence at all on the actual answers. As there are four parameters per item in the polytomous model, and their estimates are of not much use in judging the item locations, they are not reported in Table 7.1. Instead, the value of the latent variable yielding two (half of the maximum) as expected value is reported, under the symbol β^* .

Table 7.1 Parameter estimates of the 'extraversion' scale

Adjective		Continuous		Polytomous
Dutch	English	η	δ	β^*
Extravert	(Extravert)	1.932	11.224	0.610
gereserveerd ^a	(Reserved)	-1.573	12.007	-0.176
praatgraag	(Talkative)	0.024	8.584	-0.130
terughoudend ^a	(Aloof)	-1.382	12.437	-0.032
verlegen ^a	(Shy)	1.601	7.661	0.298
zwijgzaam ^a	(Taciturn)	-0.734	8.160	-0.298
introvert ^a	(Introvert)	0.132	4.943	0.189
Mean		2.5004		1.1416
SD		5.2180		0.7067

^afor these adjectives, the complementary scores were used

As is seen from the table the estimated SD for the continuous model is much larger than for the PCM model, and this is consistent with the large numbers along the horizontal axis in Fig. 7.2. The correlation between the η - and β^* -parameters is 0.85.

The latent variable estimated by the PCM will be indicated by θ_p ; the one estimated by the continuous model by θ_c . Their means and SDs are indicated by the subscripts ‘ p ’ and ‘ c ’ as well. The variable θ_p will be linearly transformed such that the transformed variable has the same mean en SD as θ_c , i.e.,

$$T(\theta_p) = B\theta_p + A.$$

It is easy to check that

$$B = \frac{\sigma_c}{\sigma_p} \text{ and } A = \mu_c - B\mu_p$$

In Fig. 7.4, the two information functions for the seven item checklist are given, for values of the latent variable brought to the same scale where the population mean is 2.5 and the SD about 5.22. The vertical dashed line indicates the average of the latent value distribution, and the thick black line indicates the range that encompasses 95% of the distribution. So, for the great majority of the population the information provided by the continuous model is larger than for the discrete PCM.

Of course the procedure we followed can be criticized: no account has been given to the standard errors of the estimates, and the data do not come from a sample that has been tested twice. So, silently, we have assumed that the distribution of the latent trait has not changed (very much) in the period that the data were collected. As in all these cases the population consists of Psychology freshman at the University of Amsterdam, it is fairly unlikely that the distribution has changed in any important way.

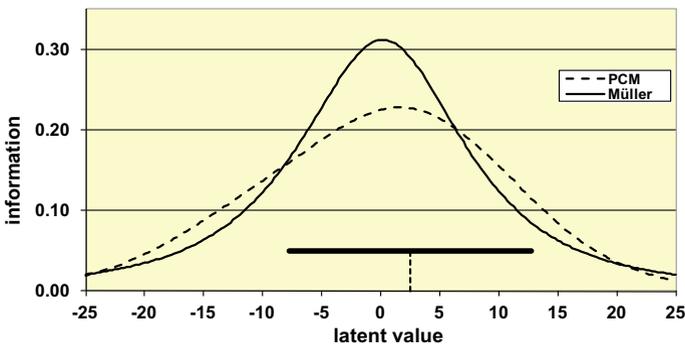


Fig. 7.4 Information functions for the PCM and for Müller’s model for the same adjective checklist

Even with all the criticism that may be put forward, it seems quite convincing that the continuous model is more informative than the PCM and therefore certainly deserves more attention than it has received thus far.

7.5 Discussion

Two IRT models for continuous responses that allow for separation between item and person parameters have been studied. Although CML estimation of the item parameters is possible, it leads to unwieldy formulae and only the case for two items has been considered for the two models. This, however, proves to be sufficient to obtain consistent estimates of the item parameters, using the theoretical contributions in the area of pseudo-likelihood estimation.

Using the technical details provided in this chapter, developing appropriate software for parameter estimation is not hard. For the extension of Müller's model, the software developed applies the Newton-Raphson procedure directly on the initial estimates and works very well, although, admittedly, it was only used on a scale with no more than 7 items. For the extension of the Rasch model to continuous responses, software has to be developed yet.

But there are more aspects in the technical realm that have to be given attention to:

1. The estimation of the individual person parameter, not only as a technical problem, but also as a real statistical one: is the maximum likelihood estimator of θ , used with known values of the item parameters, unbiased, and if it is not, can one develop an unbiased estimator along the lines of the Weighted Maximum Likelihood or Warm-estimator?
2. The set-up of goodness-of-fit tests which are informative, e.g., for the contrast between the two models discussed here, but also techniques to detect non-fitting items, all based on sound statistical reasoning are areas which deserve attention.
3. The extension of Müller's model to allow for different discriminations turned out to be rather simple, and as the example in Table 7.1 shows, worthwhile as the estimates differ considerably. Therefore, it also seems worthwhile to think about an extension of the Rasch model for continuous items that allows for different discriminations. A promising route would be to explore the possibilities of Haberman's (2007) interaction model.

There are, however, aspects in testing that cannot be solved by sophisticated techniques but which are fundamental in the discussion of the validity of the test and the conclusion its use leads to. Take the small scale on extraversion which was discussed in Sect. 7.4. If the answers are collected using a Likert scale, there is ample evidence in the literature for certain tendencies (like the tendency to avoid extreme answers) which will create irrelevant variance in the test scores. Suppose we have two respondents who are quite extravert, i.e., who would on average score at or above the middle of the five categories, but one, A, has a strong tendency to

avoid extreme answers while the other, B, has a preference for extreme answers, then on the average B will obtain a higher test score than A, while the difference cannot be unambiguously be attributed to a difference in extraversion, i.e., some of the variance of the observed scores is due to a variable which is irrelevant for the trait to be studied, and therefore forms a threat to the correct interpretation of the results, i.e., to the validity.

More in general, there are always determinants of behavior which are responsible for part of the variance in the responses, and which are a kind of a nuisance in the interpretation of the test scores, but which cannot be easily avoided. If these extra variables are associated with the format of the items then they are hard to discover if the use of this format is ubiquitous, like the multiple choice (MC) format. It is highly probable that ‘blind guessing’ as a much used model for explaining the behavior in MC test is highly unrealistic; students use all kinds of strategies to improve the result when they are not sure of the correct response, and some of these strategies will give better results than others, so that the result on an MC test is a contamination of the intended construct and the cleverness in using choice strategies.

As long as there is no variation in the format of the items, this contamination will not do much harm, as the irrelevant construct will be absorbed into the intended construct. But the risk of incorrect interpretations occurs if at some point a drastic change in the format is introduced, like switching from MC to probability measurement.

When Dirkwager was developing his system of multiple evaluation—the name he gave to probability measurement—he was aware of a serious validity threat: students can show a lack of ‘realism’ when assigning continuous responses to each of the alternatives of a multiple choice question, either by being overconfident and giving too high a weight to one of the alternatives, or by being cautious and tending to a uniform distribution over the alternatives. here is a quotation from Holmes (2002, p. 48): *Considering one item in a test, we can represent the student’s uncertainty as to the correct answer by a probability distribution $p = (p_1, p_2, \dots, p_k)$ over the set $\{1, 2, \dots, k\}$ where $0 \leq p_j$ and $\sum p_i = 1$. The student’s response can be represented by $r = (r_1, r_2, \dots, r_k)$. For a perfectly realistic student the response r is equal to p . In effect, such a student is stating: “Given my knowledge of the subject, this item is one of the many items for which my personal probability is $p = (p_1, p_2, \dots, p_k)$. For these items, answer one is correct in a proportion⁸ of p_1 , answer 2 in a proportion of p_2 etcetera.*

This means that a student is realistic if his continuous responses match his subjective probabilities, which is already elicited by the (approximate) logarithmic scoring function, but for which Holmes developed a measure of realism (see his Chap. 4) and showed that using this measure as feedback was very effective in changing the behavior of ‘unrealistic’ students, which makes the multiple evaluation approach as developed originally by Dirkwager a good and powerful instrument for formative assessment. Details of the application can be found in Holmes.

Finally, one might ask then why an IRT model is needed if good use can be made of probability measurement. There are, however aspects of formative assessment

⁸The quotation was a bit changed here as Holmes spoke of p_1 cases, clearly confusing frequencies and proportions.

which are very hard to develop at the school level. Progress assessment, for example, is an important one, but also the choice of items whose difficulty matches the level of the students. The construction of an item bank at the national level, for example, and making it available to the schools together with the necessary software, could be a task for a testing agency having the necessary psychometric and IT competences.

Appendix

To estimate the parameters in Müller's model the logarithm of the pseudo-likelihood function is maximized. The function, given in (7.30) is repeated here:

$$\sum_v \ln PL_v[\varepsilon, \eta; (r_{v1}, \dots, r_{vk})] = \sum_v \sum_{i < j} \ln f_{ij}(r_{vi}, r_{vj} | r_{vij}),$$

where v indexes the respondent. Concentrating on a single answer pattern, and dropping the index v , we have

$$f_{ij}(r_i, r_j | r_{ij}) = \frac{\sigma_{ij}^{-1} \varphi(x_{ij})}{\Phi(Z_{ij}) - \Phi(z_{ij})} \quad (7.31)$$

with the auxiliary variables

$$\begin{aligned} \sigma_{ij}^2 &= \frac{1}{2(\delta_i + \delta_j)}, \\ \mu_{ij} &= \sigma_{ij}^2(\varepsilon_i - \varepsilon_j + 2r_{ij}\delta_j), \\ x_{ij} &= (r_i - \mu_{ij})/\sigma_{ij}, \\ Z_{ij} &= (M_{ij} - \mu_{ij})/\sigma_{ij}, \\ z_{ij} &= (m_{ij} - \mu_{ij})/\sigma_{ij}, \end{aligned} \quad (7.32)$$

and the two bounds, M_{ij} and m_{ij} , repeated here:

$$\begin{aligned} m_{ij} &= \max(r_{ij} - 1, 0), \\ M_{ij} &= \min(r_{ij}, 1). \end{aligned}$$

Taking the logarithm of (7.31) gives

$$\ln f_{ij}(r_i, r_j | r_{ij}) = -\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln(\sigma_{ij}^2) - \frac{1}{2} x_{ij}^2 - \ln[\Phi(Z_{ij}) - \Phi(z_{ij})]. \quad (7.33)$$

To write down the expressions for the partial derivatives w.r.t. the ε - and δ -parameters, it proves useful to define a sequence of functions $A_{p_{ij}}$, ($p = 0, 1, 2, \dots$):

$$A_{pij} = \frac{Z_{ij}^p \varphi(Z_{ij}) - z_{ij}^p \varphi(z_{ij})}{\Phi(Z_{ij}) - \Phi(z_{ij})}. \quad (7.34)$$

Using the partial derivatives of the auxiliary variables (7.31) and the functions defined by (7.33) the first partial derivatives of (7.32) are given by

$$\begin{aligned} \frac{\partial \ln(f_{ij})}{\partial \varepsilon_i} &= r_i - \mu_{ij} + \sigma_{ij} A_{0ij}, \\ \frac{\partial \ln(f_{ij})}{\partial \varepsilon_j} &= r_j - r_{ij} + \mu_{ij} - \sigma_{ij} A_{0ij}, \\ \frac{\partial \ln(f_{ij})}{\partial \delta_i} &= r_i^2 + \mu_{ij}^2 + \sigma_{ij}^2 (1 - A_{1ij}) - 2\mu_{ij} \sigma_{ij} A_{0ij}, \\ \frac{\partial \ln(f_{ij})}{\partial \delta_j} &= -r_j^2 + (r_{ij} - \mu_{ij})^2 + \sigma_{ij}^2 (1 - A_{1ij}) + 2(r_{ij} - \mu_{ij}) \sigma_{ij} A_{0ij}. \end{aligned} \quad (7.35)$$

Notice that $\frac{\partial \ln(f_{ij})}{\partial \varepsilon_j} = -\frac{\partial \ln(f_{ij})}{\partial \varepsilon_i}$ and that

$$\frac{\partial \ln(f_{ij})}{\partial \delta_j} = \frac{\partial \ln(f_{ij})}{\partial \delta_i} + 2r_{ij} \frac{\partial \ln(f_{ij})}{\partial \varepsilon_i}. \quad (7.36)$$

For the second derivatives, it turns out that we only need three different expressions; these are given next, but we leave out the double subscript ‘*ij*’ from the right-hand sides.

$$\frac{\partial^2 \ln(f_{ij})}{\partial \varepsilon_i^2} = -\sigma^2 [1 - A_1 - A_0^2], \quad (7.36a)$$

$$\frac{\partial^2 \ln(f_{ij})}{\partial \varepsilon_i \partial \delta_i} = 2\mu\sigma^2 [1 - A_1 - A_0^2] - \sigma^3 [A_0 + A_2 + A_0 A_1], \quad (7.36b)$$

$$\begin{aligned} \frac{\partial^2 \ln(f_{ij})}{\partial \delta_i^2} &= -\sigma^2 [(1 - A_1)(2 + A_1) - A_3] + 4\mu\sigma^3 [A_0 + A_2 + A_0 A_1] \\ &\quad - 4\mu^2\sigma^2 [1 - A_1 - A_0^2]. \end{aligned} \quad (7.36c)$$

It is easily verified that

$$\frac{\partial^2 \ln(f_{ij})}{\partial \varepsilon_i^2} = \frac{\partial^2 \ln(f_{ij})}{\partial \varepsilon_j^2} = -\frac{\partial^2 \ln(f_{ij})}{\partial \varepsilon_i \partial \varepsilon_j}$$

and using (7.35), it turns out that we only need the three expressions (7.35a), (7.35b) and (7.35c) to define in a simple way the matrix of second partial derivatives. Using the symbols ‘*a*’, ‘*b*’ and ‘*c*’ to denote the value of the right-hand members of (7.35a), (7.35b) and (7.35c), respectively, one obtains the matrix of second partial derivatives as displayed in Table 7.2.

Table 7.2 Symbolic representation of the matrix of second derivatives (r means r_{ij})

	ε_i	ε_j	δ_i	δ_j
ε_i	a	$-a$	b	$b + 2ra$
ε_j	$-a$	a	$-b$	$-b - 2ra$
δ_i	b	$-b$	c	$c + 2rb$
δ_j	$b + 2ra$	$-b - 2ra$	$c + 2rb$	$c + 4rb + 4r^2a$

In the applications that were run for this chapter (see Sect. 7.4), simple initial values for the parameters were computed and immediately used in a Newton-Raphson procedure. The initial values were

$$\delta_i^{[0]} = \frac{0.1}{Var(R_i)} \text{ and } \varepsilon_i^{[0]} = 2\delta_i^{[0]} \bar{R}_i$$

where \bar{R}_i and $Var(R_i)$ denote the average and the variance of the observed responses to item i , respectively. To make the model identified, one of the ε -parameters can be fixed to an arbitrary value. To avoid negative estimates of the δ -parameters, it is advisable to reparametrize the model for estimation purposes and to use $\ln(\delta_i)$ instead of δ_i itself.

References

- Arnold, B. C., & Strauss, D. (1988). *Pseudolikelihood estimation* (Technical Report 164). Riverside: University of California.
- Arnold, B. C., & Strauss, D. (1991). *Pseudolikelihood estimation: Some examples Sankhya*, 53, 233–243.
- Albert, J. H. (1992). Bayesian estimation of normal ogive item response functions using Gibbs sampling. *Journal of Educational Statistics*, 17, 251–269.
- Andrich, D. (1982). An extension of the Rasch model for ratings providing both location and dispersion parameters. *Psychometrika*, 47, 105–113.
- Bejar, I. I. (1977). An application of the continuous response model to personality measurement. *Applied Psychological Measurement*, 1, 509–521.
- Cox, D. R., & Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91, 729–737.
- De Finetti, B. (1965). Methods of discriminating levels of partial knowledge concerning a test item. *British Journal of Mathematical and Statistical Psychology*, 54, 87–123.
- De Finetti, B. (1970). Logical foundations and the measurement of subjective probabilities. *Acta Psychologica*, 34, 129–145.
- Dirkzwager, A. (1997). *A bayesian testing paradigm: Multiple evaluation, a feasible alternative for multiple choice* (Unpublished Report, to be found in Dirkzwager, 2001).
- Dirkzwager, A. (2001). *TestBet, learning by testing according to the multiple evaluation paradigm: Program, manual, founding articles* (CD-ROM). ISBN:90-806315-2-3.
- Feller, W. (1957). *An introduction to probability theory and its applications* (Vol. 1) New York: Wiley.

- Ferrando, P. J. (2002). Theoretical and empirical comparisons between two models for continuous item responses. *Multivariate Behavioral Research*, 37, 521–542.
- Furley, D. J. (1978). *Statistical methods in biological assay*. London: Charles Griffin and Co.
- Fischer, G. H. (1974). *Einführung in die Theorie Psychologischer 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.
- Gough, H. G., & Heilbrun, A. B. Jr. (1983). *The adjective checklist manual*. Palo Alto, CA: Consulting Psychologists Press.
- Haberman, S. J. (2007). The interaction model. In M. von Davier & C. C. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 201–216). New York: Springer.
- Hendriks, C., Meiland, F., Bakker, M., & Loos, I. (1985). *Eenzaamheid en Persoonlijkheidskenmerken [Loneliness and Personality Traits]* (Internal publication). Amsterdam: Universiteit van Amsterdam, Faculteit der Psychologie.
- Holmes, P. (2002). *Multiple evaluation versus multiple choice as testing paradigm* (Doctoral Thesis). Enschede: University of Twente. Downloaded from: <https://ris.utwente.nl/ws/portalfiles/portal/6073340/t0000017.pdf>.
- Johnson, N. L., & Kotz, S. (1970). *Distributions in statistics. Vol. 2: Continuous univariate distributions-1*, (Chap. 13). New York: Wiley.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading: Addison-Wesley.
- Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149–174.
- Mellenbergh, G. J. (1994). A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research*, 29, 223–236.
- Müller, H. (1987). A Rasch model for continuous ratings. *Psychometrika*, 52, 165–181.
- Muraki, E. (1997). A generalized partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 153–168). New York: Springer.
- Noel, Y. (2014). A beta unfolding model for continuous bounded responses. *Psychometrika*, 79, 647–674.
- Noel, Y. (2017). *Item response models for continuous bounded responses* (Doctoral Dissertation). Rennes: Université de Bretagne, Rennes 2.
- Noel, Y., & Dauvier, B. (2007). A beta item response model for continuous bounded responses. *Applied Psychological Measurement*, 31, 47–73.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.
- Rippey, R. M. (1970). A comparison of five different scoring functions for confidence tests. *Journal of Educational Measurement*, 7, 165–170.
- Roby, T. B. (1965). *Belief states: A preliminary empirical study* (Report ESD-TDR-64-238). Bedford, MA: Decision Sciences Laboratory.
- Samejima, F. (1969). Estimation of ability using a response pattern of graded scores. In *Psychometrika Monograph*, No 17.
- Samejima, F. (1973). Homogeneous case of the continuous response model. *Psychometrika*, 38, 203–219.
- Samejima, F. (1974). Normal ogive model on the continuous response level in the multidimensional latent space. *Psychometrika*, 39, 111–121.
- Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). New York: Springer.
- Shuford, E. H., Albert, A., & Massengill, H. E. (1966). Admissible probability measurement procedures. *Psychometrika*, 31, 125–145.
- Toda, M. (1963). *Measurement of subjective probability distributions* (Report ESD-TDR-63-407). Bedford, Mass: Decision Sciences Laboratory.
- Van Naerssen, R. F. (1961). A method for the measurement of subjective probability. *Acta Psychologica*, 20, 159–166.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

